JESTI

Volume V
October 5th, 2026



Journal of Ethics in Scientific and Technological Innovations

Journal of Ethics in Scientific & Technological Innovation

JESTI is an annual peer-reviewed publication by the high school students at the Garcia Center for Polymers at Engineered Interfaces at Stony Brook University Summer Program (ISSN: 2834-6807).

MISSION STATEMENT

JESTI strives to provide a space for Garcia students to discuss the ethics surrounding controversial issues they may be passionate about. The writers have performed comprehensive literature research and in-depth analysis of each issue from various perspectives on a wide variety of ethical issues relating to science and technology. JESTI is an example of excellent research that can be performed by high school students when given the opportunity and proper support. It is a tool for high school students to grow and learn through research and writing experience. The students have gained valuable research, communication, leadership, and critical thinking skills through this process. We hope the future editors of JESTI will continue to embody this sentiment and keep Dr. Ellison's mission alive by enabling future generations of students to broaden their horizons and refine their concepts of ethics.

Author Inquiries:
For inquiries, please contact Dr. Miriam Rafailovich
(miriam.rafailovich@stonybrook.edu)

Front cover by Angel Zhang
Healthcare cover by Angel Zhang
Research Ethics cover by Emily Segall
Environment & Resources cover by Janet Guan
Artificial Intelligence and Technology cover by Janet Guan

Section art by Janet Guan, Emily Segall, Angel Zhang

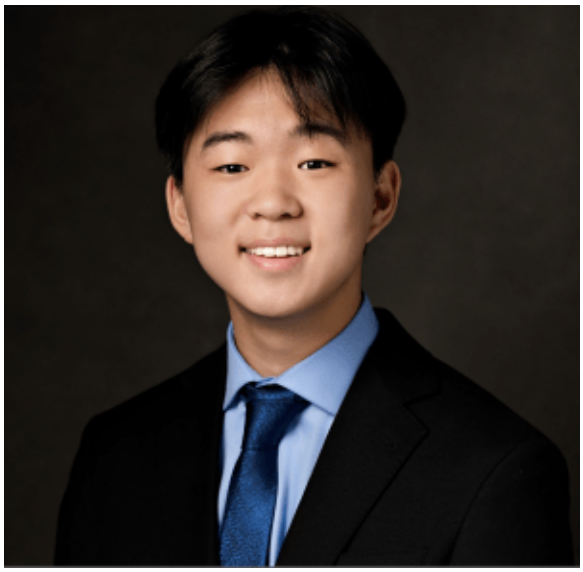
A Letter From the Editors

The editors of the Journal of Ethics in Science and Technological Innovations (JESTI) are proud to present Volume V. JESTI is a peer-reviewed, student-run journal founded in 2022 that explores today's pressing ethical issues in the fields of science, technology, and innovation. Each article offers rigorous and holistic perspectives on a host of contemporary ethical questions. Volume V was fully written, edited, and published by the high school scholars of the 2026 cohort of the Garcia Summer Scholars Research Program. Each article came from countless hours of research, discussion, analysis, critique, and revision, culminating in the sophisticated understanding of the presented ethical dilemmas. All submissions underwent an extensive peer-review process, standing testament to the intellectual scholarship that the next generation of researchers is capable of achieving.

We would like to thank our faculty advisor, Dr. Miriam Rafailovich, and our graduate student advisor, Jessica Hofflich, for their invaluable contribution to this publication. Their guidance and dedication to JESTI have created this invaluable opportunity today for high school students to develop as tomorrow's leading scientists and ethicists, and provide a space for scholars to voice their opinions as they explore wide-ranging careers. Moreover, we would like to thank our dedicated peer reviewers, whose careful evaluation ensured that submissions met a high academic and ethical standard. Lastly, we would like to thank all the student scholars who submitted their research to JESTI and acknowledge the commitment required to create such excellence. This volume of JESTI has received a record number of submissions and interest, and we are grateful and inspired by the outpouring of enthusiasm shown by this year's cohort as we continue to expand the reach and impact of JESTI.

Sincerely,
The JESTI Editorial Board

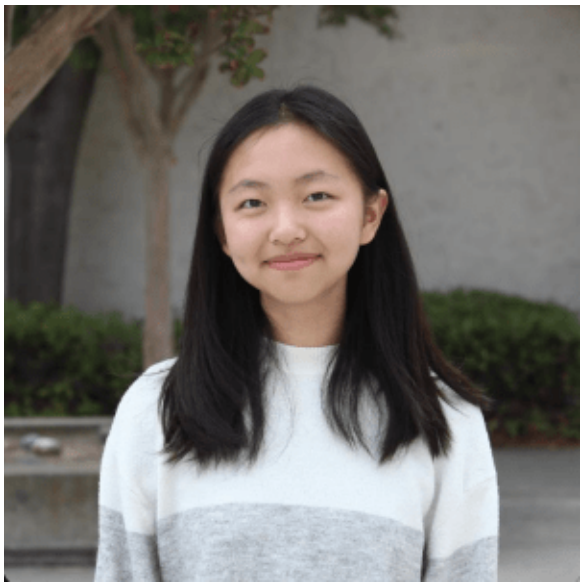
JESTI Staff Members



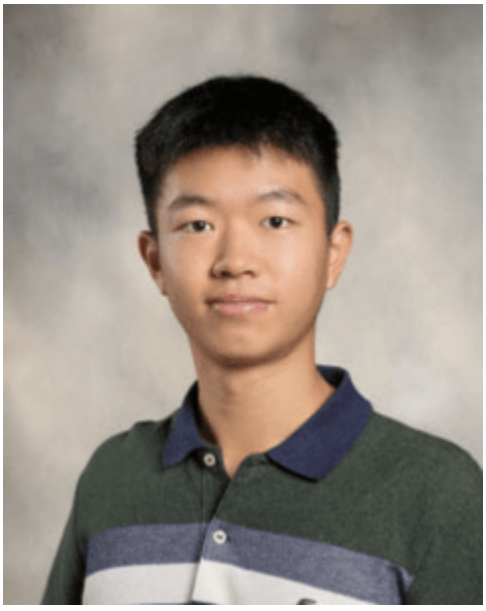
Caleb Son
Editor-in-Chief



Ria An
Editor-in-Chief



Janet Guan
Editor / Cover Artist



Leroy Jing
Editor



Marie Kouzmine
Editor



Lily Oh
Editor



Emily Segall
Cover Artist



William Jiang
Editor



Angel Zhang
Cover Artist



Jessica Hofflich
Graduate Advisor



Miriam Rafailovich
Faculty Advisor

TABLE OF CONTENTS

Title	Author(s)	Pages
Topic 1: Healthcare		
The Ethics of Disclosure: Body Dysmorphic Disorder and Cosmetic Surgery	Jacob Stern	1-11
Physician Assisted Dying: Dignified Mean to an End, or Hippocratic Hypocrisy	Max Weiner	12-17
Ethical Implications of Regenerative Medicine: Potential Societal Struggles and Solutions	Yiting Bao, Ashley He	18-34
Therapy Versus Enhancement: The Dual-Use Ethics of Emerging Technology	Marie Kouzmine	35-41
The Access Paradox: Who Gets to Benefit When Medicine Becomes a Market	Solenyi Onwueme	42-47
The Ethics of Preserving Identity During Awake Craniotomy	Angel Zhang	48-57
Editing the Self: Ethical Implications of Therapeutic Memory Modification	Lily Oh	58-66
The Limits of Consent in Adaptive Deep Brain Stimulation	Jayden Lin	67-80
Topic 2: Research Ethics		
"Research Use Only:" Mislabelling in Counterfeit Peptides in the Grey Market	Audrey Tan	81-89
Priced Before Proven: How Speculative Markets Interact with Scientific Discovery	Caleb Son	90-97
A Mammoth Undertaking: Ethical Limits on the Commercialization of De-Extinction	Lydia Choi, Bella Lin, Lelin Tsai, Andrew Wang	98-113
Uncaging the Truth: The Ethics of Using Animals for Research	Emily Segall	114-136

Assessing the Ethics of Trials That Depend on Risk: Studying People Who Inject Drugs	Luke Zhang	137-142
Kiruna, 2021: A Case Study into Moral Hazard and a Test for Moral Hazard Objections to Research	Sean Park, Luke Zhang	143-157
Topic 3: Environment & Resources		
Seeding the Oceans: The Divide Between Marine Conservation and Progressive Science	Alex Fu, Leroy Jing, Roy Lin	158-164
The Ethics of Engineering for Performance versus Sustainability in Sports Equipment	Grace Zeng	165-171
A Review of the Mechanics, Efficacy, and Ethical Concerns Associated with Enhanced Rock Weathering	Xuanming Liu	172-182
The Space Race's Second Lap! An ethical discussion regarding the modern space revitalization	Gavin Finkel	183-199
What It Takes to Sell: An Economic and Ethical Evaluation of Planned Obsolescence in Consumer Electronics	Melody Dai, Janet Guan	200-210
Digging the Hidden Toxicity: The Ethical Analysis of PFAS in Daily Consumer Goods and Drinking Water	Megan Li	211-222
The Ethics of Incorporating Nanotechnology into Food Packaging and Food Storage Materials	Taylor Maguire, Ariana Lye	223-234
Ethical and Sustainability Challenges in Modern Agricultural Production	Justin (Jiahong) Ren	235-244
Domestic Regulation and Global Responsibility: The Case of Desflurane	William Neice, William Jiang, Joel Pae, Anish Peri	245-261
Topic 4: Artificial Intelligence & Technology		
An Ethical Three-Layer	Ria Sunwoo An	262-277

Framework for Evaluating AI Responses to Mental Health Crises		
Analysis of Humanitarian Effects in the Production of Critical Materials	Henry Huang	278-287
Human-Centered Labor in the Age of AI: A Framework for Preserving Social and Emotional Work Structures	Isaac Sedighi	288-293
The Limits of Algorithmic Authority: When Is Physician Override of Diagnostic AI Ethically Justified?	Steve Samuel	294-313
The Impossibility of a Neutral Optimizer	Konstin Lee	314-329

Journal of Ethics in Scientific and Technological Innovations

The Ethics of Disclosure: Body Dysmorphic Disorder and Cosmetic Surgery

Jacob Stern[1]
[1]*Davis Renov Stahler Yeshiva High School for Boys, New York, NY 11598*

Abstract

Body dysmorphic disorder (BDD) is a psychiatric condition affecting approximately 1 to 2 percent of the population.[1,2] Those afflicted with BDD become preoccupied with a perceived flaw that others cannot see. In patients presenting for cosmetic surgery, approximately 7 to 13 percent report symptoms meeting criteria for BDD.[2,3] Several published studies indicate that surgery rarely relieves the disorder and sometimes exacerbates it.[2,4] This paper examines the ethical question of whether a surgeon who knows that a patient has been diagnosed with BDD should disclose published findings involving other patients with BDD who have received cosmetic surgery. The Code of Ethics of the American Society of Plastic Surgeons requires the disclosure of potential adverse outcomes with their probability and severity in order to secure informed consent.[5] This general principle is well settled. Since the Code of Ethics does not address BDD specifically, the principle should also be applied to BDD in a manner that satisfies the surgeon's obligation to obtain informed consent and preserves the autonomy of a patient.

1. Introduction

Body dysmorphic disorder (BDD) is defined in the DSM-5 as a preoccupation with one or more perceived defects in physical appearance that are "not observable or appear slight to others."[6] The manual sets three further criteria: (1) repetitive behaviors or mental acts performed at some point during the course of the disorder in response to the appearance concern, such as mirror checking, grooming, or comparing oneself with others; (2) clinically significant distress or impairment in social, occupational, or other important functioning; and (3) a preoccupation not better explained by concerns about body fat or weight in a person meeting criteria for an eating disorder.[1,6]

The disorder entered the manual in 1980, in the third edition, where it was described as "an atypical somatoform disorder," and was given formal diagnostic status and its present name in 1987.[1,2] The somatoform category comprised conditions marked by physical symptoms suggesting a physical disorder for which no organic cause could be demonstrated. BDD was listed among them, alongside hypochondriasis and conversion disorder.[7]

The somatoform classification did not go unchallenged, and surgeons had the challenge in their own literature before the manual acted on it. Writing in Plastic and Reconstructive Surgery in 2006, Crerand, Franklin, and Sarwer reported that the somatoform classification had been criticized, and that some researchers held the condition should instead be considered "an obsessive compulsive spectrum disorder."[2] They reported a neuroimaging study in patients with the disorder finding abnormalities similar to those seen in obsessive compulsive spectrum conditions.[2]

The manual was revised in 2013, when the fifth edition placed the disorder within the obsessive-compulsive and related disorders.[1,6] By 2017, the classification was settled. A multicenter study in JAMA Facial Plastic Surgery described the condition simply as "an obsessive compulsive-related disorder."[3]

The reclassification bears on whether surgery is likely to relieve the disorder. Crerand, Franklin, and Sarwer report that cosmetic treatment typically produces no change in the appearance concerns of patients with the disorder, and in some cases worsens symptoms.[2] Preoccupation with a single feature is also uncommon, as most patients focus on three or four over the course of the illness.[1] If the preoccupation is sustained by a compulsive process rather than by the feature itself, as the obsessive-compulsive classification indicates,

then correcting the feature would not be expected to resolve it.

The disorder is more common among people seeking cosmetic surgery, though how much more depends on how it is measured. Two American cosmetic surgery studies that are frequently cited report rates of 7 and 8 percent, both assessed by self-report questionnaire rather than clinical interview.[2] A larger and more recent American study, prospective and conducted across three practice settings, found 13.1 percent of patients presenting for cosmetic procedures reporting symptoms meeting criteria on a validated instrument, and estimated that 9.9 percent would be confirmed on structured clinical interview.[3] Two recent reviews pooling data across dozens of studies report higher figures: 21.6 percent across plastic surgery studies and 14.0 percent across American studies;[8] and 24 percent among plastic surgery patients and 12 percent across North America.[9]

2. The Question

This paper examines what a surgeon should disclose to a patient known by the surgeon to have been diagnosed with – and not merely suspected of having – BDD before conducting cosmetic surgery on that patient. The question is whether a surgeon should tell a patient what has been published about how patients with a BDD diagnosis fare after cosmetic surgery for the patient's consent to the procedure to be informed.

3. What the Outcome Data Shows

Cosmetic surgery outcomes in patients with BDD have been studied for over two decades. Reported results vary considerably across studies, several of which have been reviewed and are discussed herein.

Crerand, Phillips, Menard, and Fay examined the nonpsychiatric treatment histories of 200 individuals with BDD, recruited substantially through mental health professionals.[4] Thirty-nine of these subjects received a total of 77 surgical procedures.[4] Symptoms of the disorder improved following 1 of the 77 procedures, or 1.3 percent.[4] Symptoms of the disorder were unchanged following completion of 69 of these procedures, or 89.6 percent.[4] BDD symptoms deteriorated further following 7 of the procedures or 9.1 percent.[4] In connection with these results, the appearance of the treated body part was reported to have improved after 27 of the same procedures or 35.1 percent.[4] The authors called the contrast striking, observing that because patients may report improvements in appearance after receiving surgery, surgeons may be unaware that symptoms of BDD rarely improve.[4] In other words, notwithstanding the purported success of a cosmetic surgery procedure positively impacting one's physical appearance, the psychological impact and symptoms of BDD largely do not improve.

Crerand, Menard, and Phillips returned to the same group several years later, isolating surgical and minimally invasive procedures.[10] The appearance of the treated body part improved after roughly a third of procedures.[10] Preoccupation with that particular part decreased following 25.3 percent of procedures.[10] Overall symptoms of the disorder – preoccupation together with distress or functional impairment, counting body parts other than the one treated – improved following 2.3 percent of the procedures.[10] Among procedures producing no lasting improvement, 43.4 percent produced temporary improvement in appearance or preoccupation.[10] The same study found that physicians refused requested surgical and minimally invasive procedures less often than they refused other cosmetic treatments.[10] Furthermore, the analysis found that cost, rather than refusal, was the most common reason a sought procedure went unperformed, at 29.5 percent versus 26.2 percent.[10]

Three features appear to frame the conclusions that may be drawn from these analyses. First, the unit of analysis is procedures instead of patients. Accordingly, a subject who underwent four operations contributes four distinct observations instead of one. Next, the data from the study are retrospective and rely on a patient's recall after the procedure. Finally, two-thirds of subjects were in psychiatric treatment when assessed, which the authors acknowledge could have biased the sample

toward patients for whom nonpsychiatric treatment had already failed.[4]

Picavet and colleagues addressed each of those limitations in a study published in Plastic and Reconstructive Surgery in 2013.[11] One hundred sixty-six adults presenting for a rhinoplasty procedure were assessed before surgery on the BDD-modified Yale-Brown Obsessive Compulsive Scale (BDD-YBOCS) compared against a control group, and followed for a year using validated satisfaction and quality-of-life instruments.[11] Contrasted with the study performed by Crerand, Picavet's study was prospective; the patients were recruited in a surgical clinic rather than a psychiatric one, and symptom severity was recorded before any outcome was determined.

Notably, a patient's severity of symptoms prior to surgery inversely correlated with the patient's satisfaction after the surgery at three months and again at twelve, and correlated positively with disability and with appearance-related disruption of everyday life. When dividing patients assessed on the BDD-YBOCS at a score of 16, those with moderate to severe symptoms rated their postoperative nasal appearance 5.41 against 7.72 for patients with no or mild symptoms. Likewise, when evaluating patients on the Rhinoplasty Outcome Evaluation, which is a validated measure of satisfaction with rhinoplasty results, those patients with moderate to severe symptoms scored 14.78 against those with no or mild symptoms who scored 19.00. Moreover, quality of life in the moderate-to-severe group remained significantly lower one year after surgery.Picavet also compared patient self-assessments of nasal shape versus assessments made by surgeons. Notably, there was no correlation between the two groups assessing nasal shape.[11] The conclusion that may be drawn is that the opinion of a trained observer cannot reliably predict how the patient would view the outcome.

In a study evaluating 31 women who had rhinoplasty surgery and exhibited mild to moderate symptoms of BDD, Felix and colleagues report considerably more favorable results.[12] Notably, 21 percent of the women seeking rhinoplasty who exhibited severe BDD symptoms were excluded from the study and referred for psychiatric treatment.[12] In these patients, cosmetic surgery was determined to be contraindicated, meaning the operation is not expected to produce the benefit it is meant to produce, and may make that patient worse. In the study, severity of BDD was assessed, in part, by measuring scores on the Body Dysmorphic Disorder Examination (BDDE), which measures patients on a scale of 0 to 168. In the BDDE, 66 is the score at which patients are diagnosed as having great dissatisfaction with physical appearance. In the study, patients averaged 107.2 prior to surgery. Severity was classified by the presence or absence of overt avoidant behavior in addition to a BDDE score of 66 or higher.[12] The examination was performed by a psychologist through an interview averaging forty-five minutes, and the operating surgeon was unaware of the severity of BDD in each patient on whom the surgeon operated. One year after surgery, 25 of the patients, or 81 percent, were in complete remission and no longer met the criteria for BDD. Likewise, 28 of the patients or 90 percent were satisfied with the surgical result.[12]

In a letter to the same journal, Crerand and Phillips challenged the findings of Felix and colleagues. The letter noted that 17 of the 31 patients were assessed by both laypersons and surgeons to have a moderate to marked nasal deformity. This determination places these patients outside the diagnostic criteria for BDD, which requires that the perceived defect be nonexistent or slight. Furthermore, all 31 patients remained preoccupied with their perceived defect for at least an hour a day one year after surgery. Crerand and Phillips regard the ongoing preoccupation as compatible with BDD rather than remission.[13]

In a review of 11 studies—6 of which were retrospective and involved looking back at patients on whom surgery was already performed and 5 of which were prospective and involved patients who were enrolled in the study before surgery was performed—Bowyer and colleagues conclude that cosmetic treatment is likely ineffectual for most individuals who have

BDD. Notwithstanding this general conclusion, the review leaves open the possibility that certain individuals with mild BDD and localized appearance concerns may benefit from surgery. The review also criticizes retrospective studies since they rely on patient recall and originate in psychiatric settings, which have a bias toward surgical failures and patients who appear with more severe forms of BDD.[14]

In another retrospective study involving 728 patients at an oculofacial plastic surgery practice, Woolley and Perry screened for BDD using the Dysmorphic Concern Questionnaire. The Dysmorphic Concern Questionnaire consists of seven self-reported questions that measure patient concern about physical appearance on a scale from 0 to 3, with a total maximum score of 21. A Dysmorphic Concern Questionnaire cutoff point of greater than or equal to 9 has been validated as a clinical screening tool for BDD. Of the 728 patients, 50 of them, or 6.9 percent, scored at or above the cutoff. Each was compared against three patients from the same month who scored below it. The patients who screened positive experienced significantly higher rates of complications after surgery, reported higher pain after surgery, and required more reoperations than the control group.[15] Unlike the measures used in the studies above, these outcomes were recorded in the patient's chart rather than reported by the patient.

Veale surveyed 25 patients diagnosed with BDD who underwent a total of 38 cosmetic surgeries and 8 minimally invasive procedures. After half of the combined procedures were performed, a patient's preoccupation transferred to a different area of the body. For example, only three of 17 patients who had a rhinoplasty sought a revision of that surgery, while the remaining 14 returned for an entirely different procedure once the patient's preoccupation transferred to a different area of the body. Veale observed that the study had selection bias in favor of treatment failures, since patients were identified during psychiatric assessment.[16]

Taken together, these studies yield different forecasts for different patients who are diagnosed with BDD. Outcomes vary with how severely patients are affected. Moreover, studies reporting favorable results excluded patients who were most severely affected before surgery was offered. Enhancement of the treated physical feature is a poor guide to whether BDD improves, and any improvement that occurs is frequently only temporary.[10,16]

During consultations, this information is not available to a patient with BDD. A patient is aware of how she feels about how she looks. However, that same patient is unaware that her surgeon's view of the outcome of cosmetic surgery may not predict her own. Likewise, the patient is unaware that patient satisfaction falls as BDD symptoms rise. Nor does she know that the treated physical feature may be fixed while the underlying BDD is not. A patient's surgeon can find this information in medical journals. However, the patient is unlikely to encounter any of it on her own.

4. What Is Informed Consent and What Does It Require?

The American Medical Association (AMA) states: "Informed consent to medical treatment is fundamental in both ethics and law. Patients have the right to receive information and ask questions about recommended treatments so that they can make well-considered decisions about care. Successful communication in the patient-physician relationship fosters trust and supports shared decision making."[17]

Informed consent occurs when communication between a patient and physician results in the patient's authorization or agreement to undergo a specific medical intervention. In seeking a patient's informed consent, a physician should provide information accurately about "the burdens, risks, and expected benefits of all options, including forgoing treatment." After providing this information, the physician should document the informed consent conversation and the patient's decision in the medical record.[17]

The American Society of Plastic Surgeons (ASPS) states the same principle in its own Code of Ethics. The Code of Ethics concludes with a Statement of Principle of

Informed Consent, which describes the doctor-patient relationship as one of shared decision-making. This statement also affirms that shared decision-making "is at the heart of the doctor-patient relationship and is based on the ethical principles of respect for individual autonomy and dignity."[5] The Statement of Principle of Informed Consent further explains that for any surgical operation or treatment, "relevant information must be provided, discussed, and understood by the patient."[5]

The Statement then lists several categories of relevant information, including the nature of the surgery or treatment, expected benefits, consequences and side effects of the operation, potential risks and adverse outcomes with their probability and severity, and anticipated outcome. Section 1 of the Code also requires doctors to disclose new medical advances to patients.[5] This duty contemplates disclosure that a surgeon initiates rather than one he only provides upon request by a patient.

The results of studies, such as those described above in Section 3, fall within the category identified in the Code of Ethics as "potential risks and adverse outcomes with their probability and severity."[5] These studies report on procedures and whether BDD improved, remained unchanged, or worsened. The studies also report when a patient's preoccupation transferred to another part of the body. In addition, the studies report on rates of reoperation and the relationship between the severity of symptoms and satisfaction. A surgeon who does not provide this information to a patient known to have BDD has not provided information that the Code of Ethics requires that surgeon to provide.[5]

Neither the ASPS Code of Ethics nor AMA Opinion 2.1.1 specifically mentions BDD. However, the principles governing both the ASPS and AMA direct physicians to the same place. Both require physicians to provide accurate information bearing on the probability and severity of adverse outcomes to their patients. Likewise, the level of information that is provided must be sufficient to weigh the patient's options, including the option of not undergoing surgery. For a patient known to have BDD, the results of the studies described in Section 3 are that information.

5. How Consent Forms Address Mental Health

The ASPS provides its members with ASPS Informed Consent Resource forms for a number of surgical procedures. According to the ASPS website, these forms "are a vital tool that protect you and your practice by educating patients of the potential risks and benefits of their specific surgical procedure."[18]

The ASPS publishes a list of the forms contained in the current 2024 edition of the Informed Consent Resource. There are over 100 forms that are organized into nine categories, covering procedures from abdominoplasty to rhinoplasty, along with forms that are intended for general use.[19] The general use forms are titled General Risks of Surgery, Instructions for Using Informed Consents, COVID-19 Risk, Hepatitis and HIV Test, Removal of Jewelry and Accessories, Risks of Alcohol on Surgery, and Risks of Marijuana on Surgery. Based on the table of forms, none of the forms in the 2024 Informed Consent Resource specifically addresses mental health or BDD.[19]
This section examines a number of forms from 2020, such as the General Risks of Surgery form,[20] the Preoperative Patient Information form,[21] and consent forms for individual procedures including abdominoplasty,[22] rhinoplasty,[23] and partial facelift.[24] None of the forms examined mentions BDD.[20,21,22,23,24]

The General Risks of Surgery form directs a patient to inform the doctor as to whether the patient has a history of depression or mental health disorders.[20] The form also provides that while many people are happy after surgery, it is impossible to predict what effect surgery may have on a patient's mental health.[20] When considered against a patient known to have BDD, this last statement is difficult to maintain. Indeed, when reviewing other studies, Crerand and colleagues observed in Plastic and Reconstructive Surgery in 2006 that cosmetic treatment often does not remedy a BDD patient's concerns regarding appearance. In some cases, BDD symptoms even worsened.[2] Furthermore,

the studies described above in Section 3 report an array of outcomes and sometimes reach contrasting findings. Nevertheless, these studies address the very question the General Risks of Surgery form asserts is "impossible to predict."

The General Risks of Surgery form states that there is no guarantee of the final results of the surgery. The form also advises that nobody is perfectly symmetrical or even and that many issues with unevenness cannot be fully corrected with surgery. Patients are also advised that some patients never reach their desired goals, that the issue for which surgery was performed may come back, and that patients may choose additional surgery to get a result that the patient is happier with.[20] These general statements, which essentially advise patients that results are not guaranteed, serve as a disclaimer rather than a patient disclosure. The Code of Ethics directs surgeons to advise patients on the probability and severity of adverse outcomes.[5] Provisions in a General Risks of Surgery form that are intended to apply broadly to patients do not supply the probability and severity the Code of Ethics contemplates for a patient known to have BDD. Moreover, Crerand and colleagues reported that the appearance of the treated body part improved after roughly a third of procedures but that overall symptoms of BDD improved after only 2.3 percent of the procedures.[10] Applied here, a surgeon who advises a patient that surgery often helps may be describing the first figure, which refers to the appearance of the treated body part. However, a patient who asks whether surgery will help may really be asking about improvement of BDD symptoms.

The General Risks of Surgery form addresses several other risk factors. How these risk factors are presented to patients is also worthy of consideration. For example, smoking receives its own section, comprising four attestations and instructions to refrain from smoking for at least six weeks prior to surgery. The form also provides authorization for the surgeon to move a scheduled operation to a later date and to test for nicotine via a urine or blood test immediately before surgery. Similarly, sleep apnea receives its own section with a seven-symptom checklist. Deep vein thrombosis and pulmonary embolism have a dedicated section with a ten-item checklist and an attestation.[20]

The Preoperative Patient Information form directs the patient to list drug allergies, prescription medications, and nonprescription medications, supplements, vitamins, and herbs.[21] It also asks the patient to attest by checkbox to smoking status and secondhand smoke exposure.[21] Nothing in the form is directed to psychiatric history.[21]

6. Disclosure Rather Than Refusing to Operate

There are opinions which suggest refusing to operate as a suitable—and even preferred—alternative to disclosure.[25,26] Hostiuc and colleagues advocate that a surgeon who suspects a patient has BDD should abstain from performing surgery and refer that patient for psychological evaluation. In that context, there are but three courses of action: to operate, to refuse to operate, and to refer for evaluation for BDD. Disclosure is not offered as an alternative course of action.[25]

Similarly, Lai and colleagues advocate a practice that refuses to operate on patients with BDD. After reviewing records of 817 patients who wanted cosmetic surgery over a period of three years, 63 patients were identified as having BDD. Of those patients, 54 were identified prior to undergoing surgery and were advised against having surgery. Their conclusion was to avoid performing cosmetic surgery on patients with BDD.[26]

Although some sources advocate for refusal, others undercut the effect that refusal is intended to have on BDD patients. Refusal is intended to prevent cosmetic surgery that is unlikely to help and direct a BDD patient to seek psychological help. Some studies that are presented below indicate that neither result reliably occurs.

Veale described nine patients with BDD who performed "DIY cosmetic surgery," either in desperation at being turned down for cosmetic surgery or because they could not afford it. Veale provides several examples of what kind of surgery was performed by the patients:

A man preoccupied by his skin who believed that it was too 'loose.' He used a staple gun on both sides of his face to try to keep his skin taut. The staples fell out after 10 minutes and he narrowly missed damaging his facial nerve.[16]

A woman preoccupied by her skin and the shape of her face. She filed down her teeth in order to alter the appearance of her jaw-line.[16]

A man preoccupied by his facial skin said he had used sandpaper as a form of dermabrasion to remove scars and to lighten his skin.[16]

All nine were dissatisfied with the results of their own attempts to perform DIY cosmetic surgery.[16] For these patients, being unable to obtain the desired cosmetic surgeries did not result in treatment from a psychological professional. Rather, it resulted in attempts by the BDD patients to perform cosmetic surgeries on themselves.

Spriggs and Gillam observe that the harm to a patient that results from refusing treatment is the subject of an "as-yet unasked empirical question." They note that patients with BDD have a substantially higher rate of completed suicide as compared to the general population. They further note that refusing treatment to a BDD patient is likely to be traumatic and increases the risk of suicide. After asking why not do a simple procedure to avoid trauma and not escalate suicide risk, they further propose that "the harm (distress in a person already at increased risk of suicide) done by refusing surgery may be much greater than any harm or risk from the surgery." Spriggs and Gillam also observe that other forms of management, such as medications (e.g., antipsychotic medication) and talking therapies (e.g., Cognitive Behavioural Therapy) require cooperation of the patient and are likely to be resisted by patients who believe that the problem is physical, rather than psychological.[27]

In Opinion 2.1.3 of its Code of Medical Ethics, the AMA directly addresses the act of withholding medical information from a patient. In particular, the AMA explicitly states that "[e]xcept in emergency situations in which a patient is incapable of making an informed decision, withholding information without the patient's knowledge or consent is ethically unacceptable."[28] Thus, according to the AMA, the sole basis under which a doctor may rightfully withhold information from a patient without the patient's knowledge or consent is an emergency situation where the patient is incapable of making an informed decision.

In the introduction leading up to this directive, the opinion provides that withholding information "creates a conflict between the physician's obligations to promote patient welfare and to respect patient autonomy." A surgeon who declines to inform a patient with BDD what the studies show, judging that doing so would harm the patient, has resolved that conflict in favor of his own judgment about the patient's well-being and against the patient's autonomy. By the same token, any consent to the performed procedure cannot be deemed to have been informed. Ultimately, without an affirmative request from the patient to withhold information, the AMA opinion requires disclosure.[28]

Disclosure is also the mechanism used by the medical profession in consent forms to address risks that cannot be eliminated. The General Risks of Surgery form does not instruct a surgeon to refuse to perform a surgery on a patient who smokes or on a patient with sleep apnea. In instances where there is risk, the consent forms identify risk, inform patients accordingly, and then leave the decision on whether to proceed in the hands of the patient.[20]

Nothing peculiar to BDD requires a different approach other than disclosure. A surgeon who provides a patient with published findings about BDD does not make a diagnosis of his own that the patient has BDD. When a surgeon refuses to operate on a patient with BDD, all the patient has learned is that one surgeon has refused to operate on that patient. In contrast, when a surgeon discloses the results of BDD studies to a patient who has been diagnosed with BDD, that patient learns information that the patient can evaluate and consider and take to any future consultations with other surgeons.

7. Cases of Severe BDD

The preceding sections of this paper relate to BDD patients whose BDD symptoms are mild to moderate. In contrast, for situations where BDD symptoms are severe, the sources converge against performing surgery. For example, Felix and colleagues excluded women with severe BDD symptoms and referred them for psychiatric treatment.[12] Higgins and Wysong describe patients with severe BDD as being unsuitable for cosmetic procedures. These patients are also contrasted with those with mild to moderate BDD symptoms who may benefit. Those with severe BDD symptoms were described as exhibiting avoidant behavior, impaired global functioning, social isolation, and delusional beliefs about appearance. Those with mild to moderate symptoms were described as having localized concerns and realistic expectations.[1] Similarly, Bowyer and colleagues discuss potential benefit only for patients with mild BDD and localized concerns regarding their appearance.[14]

Spriggs and Gillam do not advocate for refusal in instances of severe BDD. They argue that for an adolescent patient with BDD, the most important concern is to avoid harm, which a simple surgical procedure may accomplish. The harm Spriggs and Gillam have in mind is the distress caused by refusing surgery. Moreover, they observe that BDD patients have a substantially higher rate of completed suicide than the general population. They reason that the distress that is endured after being refused elevates the existing risk of suicide.[27]

There is a basis to doubt the connection that Spriggs and Gillam draw between refusing surgery and an increased risk of suicide. Crerand and colleagues have reported that an increased risk of suicide is a feature that is already inherent in BDD. In particular, they have observed that up to 78 percent of patients with BDD exhibit suicidal ideation. They have also reported that between 17 and 33 percent of patients attempt suicide over the course of experiencing BDD. These figures describe patients with BDD who suffer emotionally over the course of the disorder.[2] Spriggs and Gillam presume that refusing surgery increases the risk of suicide in a patient suffering from BDD. However, they offer no evidence that it does. Indeed, the ultimate question of whether the risk of suicide increases is one they identify as "unasked." Not having an answer to this question should result in a cautious approach. It should not serve as a basis for performing surgery.

In instances of severe BDD where surgery is unlikely to help a patient and may cause harm, a surgeon should decline to perform that surgery. However, refusing to perform surgery should not end the surgeon's obligation to the patient. In these instances of severe BDD, a patient is no less entitled to disclosure merely because the surgeon is also refusing to perform surgery. A surgeon who refuses the patient without explaining what the published findings demonstrate leaves the patient knowing only that one surgeon said "No."

In the studies that address the severity of BDD, it is apparent that severity is not measured in a standardized manner. For example, Felix and colleagues classified severity based on the presence or absence of overt avoidant behavior in addition to a BDDE score at 66 or above.[12] In contrast, Picavet categorized patients who scored above 16 on the BDD-YBOCS as having moderate to severe BDD.[11] In a review, Higgins and Wysong do not use any score to characterize BDD severity.[1] Accordingly, what is deemed severe BDD is not defined uniformly across the literature.

Since the severity of BDD varies across a range, any line that is drawn will separate patients who may differ just a little from one another. There will be patients standing very close to, but just short of, the line that is drawn to separate patients with severe BDD from mild to moderate BDD. For those patients who stand just short of that line, surgery remains available. These are the patients for whom the published findings are least favorable. And it is for these very patients that disclosure matters the most.

8. What Disclosure May Look Like to a Patient with BDD

The ASPS publishes forms addressing several risk factors for patients to review and

consider prior to surgery.[19] Based on the table of forms, none of the forms specifically addresses BDD.[19] The following disclosure form is influenced by the studies presented in Section 3 and is offered as an illustrative example of a disclosure form to patients with BDD. It is not a form published by the ASPS.

8.1 Informed Consent: Body Dysmorphic Disorder and Cosmetic Surgery

Instructions. This informed consent document will help you learn about cosmetic surgery on patients who have been diagnosed with body dysmorphic disorder. Please read this entire document carefully. Initial at the bottom of each page indicating that you have read the content of each page. If there is something you do not understand, please ask your plastic surgeon.

General Information. A doctor has told you that you have body dysmorphic disorder, also known as "BDD." An individual with BDD is preoccupied with a flaw in the individual's appearance that other people do not notice or think is minor. Researchers have studied what happens to patients with BDD after undergoing cosmetic surgery. This document provides you with information about the findings of studies of those individuals with BDD who have received cosmetic surgery.

Treatment for BDD. Cosmetic surgery is not a treatment for BDD. Mental health professionals treat BDD. Talk with your doctor or surgeon about these treatments prior to cosmetic surgery. You may decide not to have cosmetic surgery.

Risks of Cosmetic Surgery for Patients with BDD. Studies of patients with BDD do not all report the same or consistent results. In some cases, patients with mild or moderate BDD symptoms have reported better outcomes than patients with more severe BDD symptoms. Your surgeon cannot predict which outcome you will have. It is important that you are aware of these studies before you decide to proceed with cosmetic surgery.

8.2 Specific Risks for Patients with BDD

Effect of Cosmetic Surgery on Patients with BDD. Cosmetic surgery may physically alter the part of your body on which surgery is performed. It may not change how you feel about your appearance. In one study of 77 cosmetic operations on patients with BDD, the treated body part was reported to look better after 27 of them. However, BDD symptoms reportedly improved after only 1 procedure, stayed the same after 69 procedures, and got worse after 7. A later study of the same patients found that BDD symptoms improved after about 2 procedures out of every 100. The improvement you may feel may also be temporary. Among the procedures in that study that produced no lasting improvement, about 4 out of 10 initially made patients feel better about the treated body part for a period of time before that feeling faded.

Change in Focus to a Different Body Part. After undergoing cosmetic surgery on a particular part of your body, your concern may shift to a different part of your body. In one study involving 25 patients diagnosed with BDD, 17 patients had nose surgery. Only three patients underwent a revision of that surgery, and 14 patients returned for an entirely different procedure.

Satisfaction with Results. Your view of the result of the surgery that is performed may be different from your surgeon's view. In one study, how good the result looked to trained surgeons did not predict how the patient rated that same result.

Complications. In one study, patients who scored above a threshold on a BDD questionnaire experienced more complications, more pain after surgery, and more repeat operations than other patients.

8.3 Acknowledgments

I got this information sheet on body dysmorphic disorder and cosmetic surgery.

I understand that surgery may improve the appearance of the treated body part without improving my BDD.

I understand that any improvement in my BDD may not last.

I understand that my concern may move to a different part of my body.

I understand that my surgeon cannot tell me how I will feel about the result.

I understand that not having the surgery is an option and that I can opt out of having the surgery.

All my questions have been answered.

IT HAS BEEN EXPLAINED TO ME IN A WAY THAT I UNDERSTAND:

a. WHAT RESEARCH HAS FOUND ABOUT COSMETIC SURGERY IN PATIENTS WITH BDD

b. THERE ARE TREATMENTS FOR BDD

c. I MAY CHOOSE NOT TO HAVE SURGERY

I UNDERSTAND THE EXPLANATION AND HAVE NO MORE QUESTIONS.

Patient or Person Authorized to Sign for Patient

Date/Time

Witness

Date/Time

9. Conclusion

The AMA instructs its physicians to disclose the burdens, risks, and expected benefits of procedures.[17] The ASPS Code of Ethics requires the disclosure of potential risks and adverse outcomes along with the probability and severity. This requirement is grounded in "ethical principles of respect for individual autonomy and dignity."[5] Neither mentions BDD, nor do the consent forms published by the ASPS.[19] However, the principles governing both the AMA and ASPS direct physicians to the same conclusion: disclosure.

Nothing advocated herein requires a surgeon to make a psychiatric diagnosis or to screen patients. It asks for a surgeon to disclose results of published studies to a patient known to have BDD. The preceding section offers a demonstrative example of what a disclosure from a surgeon to a patient may look like in practice. Disclosure respects the patient's autonomy by allowing the patient to decide whether to proceed with the surgical procedure after being given the same information already known by the surgeon.

10. References

[1]Higgins, Shauna, an Ashley Wysong. "Cosmetic Surgery and Body Dysmorphic Disorder – An Update." International Journal of Women's Dermatology, vol. 4, no. 1, 2018, pp. 43-48.

[2]Crerand, Canice E., et al. "Body Dysmorphic Disorder and Cosmetic Surgery." Plastic and Reconstructive Surgery, vol. 118, no. 7, 2006, pp. 167e-180e.

[3]Joseph, Andrew W., et al. "Prevalence of Body Dysmorphic Disorder and Surgeon Diagnostic Accuracy in Facial Plastic and Oculoplastic Surgery Clinics." JAMA Facial Plastic Surgery, vol. 19, no. 4, 2017, pp. 269-74.

[4]Crerand, Canice E., et al. "Nonpsychiatric Medical Treatment of Body Dysmorphic Disorder." Psychosomatics, vol. 46, no. 6, 2005, pp. 549-55.

[5]American Society of Plastic Surgeons. Code of Ethics of the American Society of Plastic Surgeons. ASPS, 2025, www.plasticsurgery.org/documents/Governance/asps-code-of-ethics.pdf.

[6]American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 5th ed., American Psychiatric Publishing, 2013.

[7]Yeung, Albert, and He Deguang. "Somatoform Disorders." Western Journal of Medicine, vol. 176, no. 4, 2002, pp. 253-56.

[8]Kaleeny, Joseph D., and Jeffrey E. Janis. "Body Dysmorphic Disorder in Aesthetic and Reconstructive Plastic Surgery – A Systematic Review and Meta-Analysis." Healthcare, vol. 12, no. 13, 2024, article 1333.

[9]Pérez-Buenfil, Luis Ángel, and Martha Alejandra Morales-Sánchez. "Prevalence of Body Dysmorphic Disorder: A Systematic Review and Meta-Analysis." Journal of Cosmetic Dermatology, vol. 24, no. 4, 2025, article e70121.

[10]Crerand, Canice E., et al. "Surgical and Minimally Invasive Cosmetic Procedures among Persons with Body Dysmorphic Disorder." Annals of Plastic Surgery, vol. 65, no. 1, 2010, pp. 11-16.

[11]Picavet, Valerie A., et al. "Preoperative Symptoms of Body Dysmorphic Disorder Determine Postoperative Satisfaction and Quality of Life in Aesthetic Rhinoplasty." Plastic and Reconstructive Surgery, vol. 131, no.

4, 2013, pp. 861-68.
[12]Felix, Gabriel A. A., et al. "Patients with Mild to Moderate Body Dysmorphic Disorder May Benefit from Rhinoplasty." Journal of Plastic, Reconstructive & Aesthetic Surgery, vol. 67, no. 5, 2014, pp. 646-54.
[13]Crerand, Canice E., and Katharine A. Phillips. Letter. Journal of Plastic, Reconstructive & Aesthetic Surgery, vol. 67, no. 12, 2014, pp. 1754-55.
[14]Bowyer, Laura, et al. "A Critical Review of Cosmetic Treatment Outcomes in Body Dysmorphic Disorder." Body Image, vol. 19, 2016, pp. 1-8.
[15]Woolley, Andrew J., and Julian D. Perry. "Body Dysmorphic Disorder: Prevalence and Outcomes in an Oculofacial Plastic Surgery Practice." American Journal of Ophthalmology, vol. 159, no. 6, 2015, pp. 1058-64.
[16]Veale, David. "Outcome of Cosmetic Surgery and 'DIY' Surgery in Patients with Body Dysmorphic Disorder." Psychiatric Bulletin, vol. 24, no. 6, 2000, pp. 218-20.
[17]American Medical Association. "Opinion 2.1.1: Informed Consent." AMA Code of Medical Ethics, American Medical Association, 2016, code-medical-ethics.ama-assn.org/ethics-opinions/informed-consent.
[18]American Society of Plastic Surgeons. "2024 Informed Consent Resource." ASPS Shop, www1.plasticsurgery.org/shopping/product.aspx?id=41570.
[19]American Society of Plastic Surgeons. 2024 Informed Consent Download: Table of Contents. ASPS, 2024, www.plasticsurgery.org/documents/medical-professionals/shop/2024-informed-consent-download.pdf.
[20]American Society of Plastic Surgeons. Informed Consent – General Risks of Surgery. ASPS, 2020.
[21]American Society of Plastic Surgeons. Preoperative Patient Information. ASPS, 2020.
[22]American Society of Plastic Surgeons. Informed Consent – Abdominoplasty Surgery. ASPS, 2020.
[23]American Society of Plastic Surgeons. Informed Consent – Rhinoplasty Surgery. ASPS, 2020.
[24]American Society of Plastic Surgeons. Informed Consent – Partial Facelift Surgery. ASPS, 2020.
[25]Hostiuc, Sorin, et al. "Ethical Challenges regarding Cosmetic Surgery in Patients with Body Dysmorphic Disorder." Healthcare, vol. 10, no. 7, 2022, article 1345.
[26]Lai, Chung-Sheng, et al. "Body Dysmorphic Disorder in Patients with Cosmetic Surgery." Kaohsiung Journal of Medical Sciences, vol. 26, no. 9, 2010, pp. 478-82.
[27]Spriggs, Merle, and Lynn Gillam. "Body Dysmorphic Disorder: Contraindication or Ethical Justification for Female Genital Cosmetic Surgery in Adolescents." Bioethics, vol. 30, no. 9, 2016, pp. 706-13.
[28]American Medical Association. "Opinion 2.1.3: Withholding Information from Patients." AMA Code of Medical Ethics, American Medical Association, 2016, code-medical-ethics.ama-assn.org/ethics-opinions/withholding-information-patients.

Physician-Assisted Dying: Dignified Means to an End, or Hippocratic Hypocrisy?

Max Weiner[1]
[1]*Salanter Akiba Riverdale High School, Bronx, NY 10471*

Abstract

Physician-assisted dying, or PAD, is a procedure in which a medical professional aids in painlessly ending the life of a patient, usually at the patient's request. This is carried out typically via a prescribed overdose of a medication, administered either by the physician or the patient themselves. Laws over who qualifies for PAD vary between various countries which allow the practice; most reserving it for those who are terminally ill, while others allow those afflicted with chronic disease to go through with the procedure as well. The ethics and implications of PAD have long been argued, those in support claiming that it provides a dignified death to those facing a painful and ugly path, those against saying that it goes against the Hippocratic oath, allows patients to commit suicide out of depression over their ailment, as opposed to the ailment itself, and can lead to a slippery slope as to what conditions a person can have to be considered for PAD. This paper will discuss the history of PAD, the nuances within the practice which are often overlooked, and the process of the procedure itself, its current legality in different places, as well as analyze the arguments for and against the ethics of the practice both from the perspectives of patients and doctors.

1. Introduction

1.1 History of PAD

The practice of and the arguments surrounding PAD have a storied history, dating back thousands of years. The practice was actually not uncommon in Ancient Greece and Ancient Rome. In fact, PAD was far more commonly used back then than it is today.[1] Opposition to the existence of the practice in ancient times is made clear by the prohibition in the Hippocratic Oath of administering lethal drugs to someone who requests them.[2] As the centuries passed, the emergence and rise of Christianity proved another enemy towards PAD. Prominent Christian philosopher St. Thomas Aquinas condemned suicide in his Summa Theologica in the 13th Century, claiming that it was so awful as to even violate God's very authority.[3] Due to the Church holding a vast amount of power in the Middle Ages, the practice became taboo. Similarly, the English, Protestant though they were, widely prohibited PAD via common law, which spread from the English isles all the way to the Americas. Even during the Renaissance, when people's lives began to be influenced by a new set of ideals, not dictated by the Church, PAD still did not prove of any interest.

It was not until the early 19th century, with the advent of morphine and chloroform as pain killers and anesthetics, that the idea of providing a painless death became once again a topic of discussion. And once two doctors, Carl Kortuum and Christian Mursinna, independently advocated for PAD, and reported their own anecdotes detailing how they killed patients out of mercy, this topic of discussion exploded into the forefront.[4] The medical world, however, was split on the issue. Samuel Butler, a proponent for what he called 'mercy killing,' was met with significant backlash from the Journal of the American Medical Association, who likened him to an executioner.[1]

The first country to change policy over PAD was Switzerland, who legalized PAS (the definition of which will be explained shortly) in 1942. This sent waves through the medical community, as well as became a catalyst for more countries legalizing the practice.[1, 2] Legalization of PAD spread all over Europe, North America, and Australia over the course of the next 80 years.

1.2 The Modern Processes

Physician-assisted dying, or PAD, is a procedure in which a medical professional aids in ending a patient's life. This is typically done via one of a few ways, the distinction between them being of import.

PAS (Physician Assisted Suicide): In PAS, a medical professional will prescribe and/or provide a lethal overdose of a painkilling medication, at which point the patient themselves must be the one to ingest the medication.[6]

In the U.S. and Canada today, PAS also goes by Medical Assistance in Dying (MAID), and thus is how it will be referred to when in the context of the two countries.[3]

Euthanasia: In Euthanasia, on the other hand, the medical professional is tasked with administering the life ending medication, most commonly intravenously.[6]

There are three types of Euthanasia, Voluntary Euthanasia (VE), when the patient requests the procedure; Involuntary Euthanasia (IVE), when the procedure is against the patient's wishes; and Non-Voluntary Euthanasia (NVE), when the patient is unable to make a decision whether or not to accept the procedure. IVE and NVE are outlawed across the world, with VE being the only acceptable form.[6]

2. Legality and Steps of PAD Across the World

Different countries in the world have very different laws surrounding PAD, even amongst those who facilitate it. Differences include who is eligible, the process they go through to get the procedure approved, and the types of PAD they have available to them.

2.1 United States

The first state to adopt legislation permitting MAID was Oregon, in 1997.[5] Since then, across the United States, 13 states, as well as Washington D.C. have legislation (or in Montana, a court case) that permits MAID. Similarities in the process between states include:

A terminal diagnosis, with a prognosis of less than 6 months; Both the primary physician and a second physician must both independently determine the patient eligible; multiple requests for PAD from the patient, both oral and written; an explanation of alternative options by the physician; at least one witness who has no financial incentive to support the death of the patient; the patient must self administer the life ending medication; and that euthanasia is not allowed under any circumstance.[7]

The differences from state to state lie mainly in the need for a mental health evaluation. That is, in most states where MAID is legal, a mental health evaluation is only needed if one or both of the physicians analyzing the patient deem it necessary, while in the states like Hawaii and New York, anyone requesting MAID must receive a mental evaluation to deem them of sound mind to make the decision.[8]

2.2 Europe

As mentioned earlier, the first country to legalize PAD was Switzerland, since then, 8 more European countries have legalized some type of PAD.[9] While there are a few smaller differences between European countries' laws on the process of PAD, such as whether a second physician is required to approve the patient for

PAD, there are two major differences between different European countries, as well as the US and some European countries. The first being the allowance of those not with a terminal diagnosis, but instead an incurable chronic illness which causes massive amounts of mental and physical pain. Out of the 9 countries, 5 of them (Netherlands, Belgium, Luxembourg, Switzerland, and Germany) do not require any prognosis, just a painful chronic disease.[5] The second major difference is the allowance of Euthanasia. Out of the 9 countries where PAD is legal, 4 countries (Spain, Luxembourg, Netherlands, and Belgium) allow not only PAS, but also Euthanasia. Belgium, in fact, only allows Euthanasia.[9]

3. Suffer the Slings and Arrows of Outrageous Fortune Or… To Die, To Sleep no More?

As was detailed in section *1.1 History of PAD*, the debate over the practice of PAD is far from new; and just as the fight itself is one that is millennia old, many arguments used today echo arguments used during the middle ages, and even arguments used during the BCEs.

3.1 In Support

The idea of a merciful killing seems ethically contradictory, especially when that killing is done by a doctor, someone sworn to protect a patient's life above all else, to not cause harm. However, when a life isn't possible to protect, is it a doctor's duty to protect that patient's life from excess agony? That is the first issue pertinent here. Those in support of the idea of merciful killing often argue that, while death is not ideal, between death and a life of excruciating pain, death is in reality the lesser of the two harms. This concept, first identified by St. Thomas Aquinas (who, ironically, opposed PAD) is known as Double Effect; where, when harm is unavoidable no matter what, it's ok to take action to prevent an immense amount of harm, even at the cost of causing a comparatively less amount of harm.[3] Even those who are religious tend to support this concept when it comes to a medication which eases someone's pain but shortens their lifespan,[10] why should this not apply to PAD? Advocates for PAD argue that whether death is a lesser harm than living through a terminal disease (or chronic illness, depending where you are), is a question which only those suffering from these diseases and their doctors can answer for them. According to many doctors who advocate (or advocated) for PAD, the patients whose misery they put them out of were in such awful and sorry states that living any longer would be akin to torture, a torture which would only get worse the longer they went on.[4] Nurses, too, have expressed their support for PAD. In 20 studies conducted in European Hospitals, wherein Nurses were asked about their thoughts and feelings on PAD, in nearly every study a vast majority of nurses either supported the legalization of the practice, was willing to take part in a patient's "care" (so to speak), or both.[11]

Supporters of PAD argue that it is within a person's natural rights to decide to go through with PAD if they so choose. Among peoples' natural rights is the right to live a life how one wills, so long as they are not causing harm to others. Accordingly, supporters argue, just as everybody has the right to choose their own life path, so too they should get to choose how they should die, if they feel like they must. People suffering from diseases which qualify them to go through PAD are on a road to a death which is not just painful, but also strips them of their dignity, something every human is born with, and deserves to keep. So if a person decides that they would rather die slightly earlier than they would otherwise, and die in a way which allows them to keep their dignity, then they are, from the moment that they are born, allowed to choose to go out that way, so says organizations

formed for the explicit purpose of supporting PAD, such as Death With Dignity.[12]

As was written earlier, in order to go through with PAD, a person must pass through a range of checks by various people, often both physical and mental, must take time to ensure that they want to go through with the operation, and (in most places) are tasked with taking the killing medication themselves, giving them the final chance to back out. This long list of required tasks in order to go through with the procedure is one that, at least according to supporters of PAD, weeds out anyone not truly supposed to do it, and just leaves those who genuinely think they'd be better off dead.

3.2 In Opposition

As was mentioned earlier, Religion had a major part in opposing PAD. Almost every major religion (examples include Christianity, Islam, Judaism, Hinduism, and Buddhism) has a concept which discourages or forbids suicide, be it done by yourself or by someone that's paid to do it.[13] As a result, even today people who are religious (unsurprisingly) tend to stand against the practice of PAD.

In the US, a prognosis of less than 6 months is required for a person to go through with PAD, while in some places in Europe, all that's needed is a painful chronic illness. People against PAD argue that the criteria for who qualifies for PAD can get less and less restrictive as years go on.[5] One second it's people who are terminal, next people with chronic disease, what if next up is people born with some sort of disability? After all, euthanizing people over unnecessary or plain racist or ableist things is not something that humans are strangers to. It is scary to think that suicide can be a recommended option given to someone by their physician when it really is not something that is necessary. Furthermore, what if things such as NVE become legalized; if people are allowed to have a DNR order, then why not open up the floor to other methods of not saving a person's life. And what if that can lead to IVE becoming legal.

In a day where medicine is advancing at a steady rate every day, how can a doctor genuinely allow their patient to choose death, instead of holding out for hope? This is another issue that people have with PAD;[11] that a doctor would abandon the duty they swore to uphold, and go directly against what they said when they took the Hippocratic Oath. It would not only be a fault on the part of the doctor, for failing to help their patient, but also a fault on the government for facilitating the death.

One final objection used by people against PAD is that people may go through with the procedure not due to the pain of their disease, but merely the depression that they are most likely going to suffer from, being in their circumstance. While there are measures in place which are meant to prevent those who are just depressed from becoming victim of what is quite literally law facilitated suicide, those measures are inevitably going to fail at times, resulting in the death of someone who did not actually need to die, someone who, despite the checks on their mental state clearing them, is in reality being affected more by their depression than their disease. After all, people do experience and express their depression in a vast variety of ways. This is something which was addressed in a 2008 study done on Oregon residents requesting MAID. The study, done on 58 residents, found that 15 subjects were suffering from a depressive disorder, which should have eliminated them from being approved for PAD. The fact that this not-insignificant number of people from the group were found to be depressed in the first place is tragic as is. It becomes far more tragic, however, as the study reveals that 3 of the 15 individuals who had a depressive disorder were approved for PAD, and ended up dying after their 'treatment'.[14]

4. Discussion and Conclusion

The issue of PAD is one that has been debated, is being debated, and will continue to be debated long into the future. The arguments outlined in this paper, both in support and in opposition to PAD, demonstrate the inherent ethical complexity of the practice, wherein also lies a root cause for the different safeguards implemented across the world. The history of PAD covered provides insight as to how thought around the process in the medical world shifted as centuries passed. The practice, while now more widely implemented than it was a few centuries ago, still raises a clash between PAD and the Hippocratic oath, which directly forbids a doctor administering lethal medication to a patient, even upon the patient's request. The concept of Double Effect provides a defense for PAD, whereby it allows doctors the justification that killing the patient humanely is causing them less pain than keeping them alive and letting them eventually die of their disease. As has been shown, amongst those very supportive of the practice are doctors, nurses, and patients suffering from debilitating, often terminal, diseases. According to them, it is causing less harm to the patient to end their lives via PAS; however, the safeguards for the practice adds a variable which opens up even more ethical dilemmas. As for the ethical arguments which surround not the practice as a whole, but the safeguards around the practice, these vary widely due to the variance of safeguards implemented in different places in the world. Whether the safeguards around who can or cannot qualify, and how to prevent those who don't qualify from slipping through the cracks are sufficient, is a discussion of which consensus can change based on the jurisdiction the discussion is being had in relation to. A study cited earlier showed that in Oregon, multiple people who, according to the statutes put in place in Oregon, should not have been approved for PAD (due to a depressive disorder which compromised their judgement) were still allowed to go through with the treatment, leading to their deaths. A jurisdiction which conducts more rigorous mental health evaluations may not have the same issues which were presented in the Oregon study. Similarly, the concerns over the government's role in the practice, and in who they allow to go through with the practice, of course change from place to place as the government's role changes as well. A major fear in those opposed to PAD in various countries, however, is that their respective country may lower the threshold of who can or cannot be approved. People from places where PAD is limited to those with terminal diseases, and those from places where PAD is allowed for those with chronic diseases, all can fear that restrictions become ever more relaxed in their respective country. Ultimately, PAD is not a practice which can be looked at with an eye focusing on just one aspect of the practice or on just one area where it is practiced. The evidence in this paper demonstrates the immense nuance that is inherent to the practice, and the many considerations one must take when discussing its ethical standing; considerations which include legality by country or state, safeguards in place, wishes of the patients and doctors in question, and the individual circumstances of patients.

5. References

[1] Emanuel, Ezekiel J, and Steven Joffe. "Assisted Suicide and Euthanasia." *Nih.Gov*, BC Decker, 2011, https://www.ncbi.nlm.nih.gov/books/NBK13381/. Accessed 26 Aug. 2026.

[2] Stigall, William. "The Hippocratic Oath." *The Linacre Quarterly*, vol. 89, no. 3, Mar. 2022, pp. 275–86, https://doi.org/10.1177/00243639211069431. Accessed 26 Aug. 2026.

[3] Brenna, Connor T. A. "Regulating Death: A Brief History of Medical Assistance in Dying." *Indian Journal of Palliative Care*, vol. 0, no.

PMC8655630, Oct. 2021, pp. 1–4, https://doi.org/10.25259/ijpc_426_20. Accessed 26 Aug. 2026.

[4] Stolberg, Michael. "Two Pioneers of Euthanasia around 1800." *The Hastings Center Report*, vol. 38, no. 3, 2008, pp. 19–22. *JSTOR*, http://www.jstor.org/stable/25165328. Accessed 26 Aug. 2026.

[5] Mroz, Sarah, et al. "Developments after the Liberalization of Laws Concerning Assisted Dying—Experience from Belgium and Other Countries." *Deutsches Ärzteblatt International*, Deutscher Arzte-Verlag GmbH, Dec. 2022, https://doi.org/10.3238/arztebl.m2022.0378. Accessed 26 Aug. 2026.

[6] Goligher, Ewan C., et al. "Physician-Assisted Suicide and Euthanasia in the ICU." *Critical Care Medicine*, vol. 45, no. 2, Feb. 2019, pp. 149–55, https://doi.org/10.1097/ccm.0000000000001818. National Library of Medicine. Accessed 26 Aug. 2026.

[7] "State Aid-in-Dying Law Differences." *Academy of Aid-in-Dying Medicine*, 10 Feb. 2026, https://www.aadm.org/courses/state-map-and-state-aid-in-dying-laws?utm. Accessed 26 Aug. 2026.

[8] "Medical Aid in Dying Law Frequently Asked Questions." *Ny.Gov*, 2026, https://www.health.ny.gov/health_care/medical_aid_in_dying/faqs.htm?utm. Accessed 26 Aug. 2026.

[9] Francés Vañó, Lucía, et al. "Comparative Legal and Bioethical Perspectives on Euthanasia and Assisted Suicide: A Systematic Review of European Frameworks." *BMC Medical Ethics*, vol. 27, no. 1, Springer Science and Business Media LLC, Mar. 2026, https://doi.org/10.1186/s12910-026-01410-w. Accessed 26 Aug. 2026.

[10] Downar, James. "Is Physician-Assisted Death in Anyone's Best Interest? Yes." *Canadian Family Physician*, vol. 61, no. 4, Apr. 2015, p. 314, https://pmc.ncbi.nlm.nih.gov/articles/PMC4396749/. Accessed 26 Aug. 2026.

[11] Nordhaug, Marita, et al. "European Nurses' Attitudes towards Assisted Death: A Scoping Review." *Nursing Ethics*, vol. 33, no. 3, SAGE Publications, Jan. 2026, pp. 899–926, https://doi.org/10.1177/09697330251403140. Accessed 26 Aug. 2026.

[12] Death with Dignity. "Death with Dignity." *Death with Dignity*, 2025, https://deathwithdignity.org/. Accessed 26 Aug. 2026.

[13] Traina, Cristina L. H. "Religious Perspectives on Assisted Suicide." *The Journal of Criminal Law and Criminology (1973-)*, vol. 88, no. 3, Northwestern University School of Law, 1998, pp. 1147–54, https://doi.org/10.2307/3491364. JSTOR. Accessed 26 Aug. 2026.

[14] Ganzini, L., et al. "Prevalence of Depression and Anxiety in Patients Requesting Physicians' Aid in Dying: Cross Sectional Survey." *BMJ*, vol. 337, no. oct07 2, BMJ, Oct. 2008, pp. a1682–82, https://doi.org/10.1136/bmj.a1682. Accessed 26 Aug. 2026.

Ethical Implications of Regenerative Medicine: Potential Struggles and Solutions

Yiting Bao[1*], Ashley He[2*]
[1]No. 2 High School of East China Normal University, Shanghai, China 201203, [2]Tomball Memorial High School, Tomball, TX 77377

Abstract

Regenerative medicine is a branch of medicine that utilizes human stem cells, animal tissue, and biomaterials to regenerate or replace human cells, tissue, or organs. Over the last few decades, technologies such as tissue engineering, 3D bioprinting, and induced pluripotent stem cells (iPSCs) have entered this blooming field of research. While this rapidly developing field holds great potential for novel clinical treatments, several ethical implications also emerge. In this review, we focus on four areas where regenerative medicine has a notable impact—clinical translations of regenerative medicine, commercial and regulatory practices, societal impacts of accessibility and inaccessibility, and human identity. We identify potential struggles by summarizing existing literature and incorporating previous case studies, while also proposing potential solutions. This paper can serve as a source of guidance for researchers, clinical workers, the government, and the public to handle and address the ethical challenges of introducing regenerative medicines.

1. Introduction

Regenerative medicine "replaces or regenerates human cells, tissue or organ, to restore or establish normal function" and has been under extensive research for the past three decades.[1] The field can potentially address transplant organ shortage, provide alternative treatments for severe burns and diabetic wounds, and even increase life expectancy. Currently, under the branch of regenerative medicine, there have been significant developments in fields including tissue-engineered skin, 3D bioprinting, stem cell research, and tissue/organ transplantation. Although these technologies are highly distinct in their specific risks and stages of development, their similar nature as novel regenerative medicine technologies , altering the human body with foreign biomaterials, prompt us to discuss their ethical implications together. Moreover, the fields are highly interrelated; for example, tissue-engineered skin often interacts with 3D bioprinting or tissue transplantation in its development.

In this review, we utilize credible databases—Pubmed, Web of Science, and Google Scholar—to select recent science and ethics papers (mostly published after 2015) that are relevant to the development of regenerative medicine technology. We focus on four areas where regenerative medicine has a notable impact and its ethical implications must be brought into attention. First, we discuss ethical issues that arise in the clinical translation stages of regenerative medicine. This is the earliest stage of medical technology development where ethical implications in animal experimentations, human trials, and clinical implementation gaps are addressed. Second, we move on to investigating commercial and regulatory practices that shape how the technology is presented to society after going through clinical translation. We discuss discrepancies in legal categorization and misinformation spread by commercial narratives. Third, we study how accessibility and inaccessibility of regenerative

medicine will impact potential moral degradation and social stratification. This is a step further after the technology enters the market and becomes relatively common. Finally, we discuss how regenerative medicine may influence human identity. This is a final step, investigating how regenerative medicine's role in society may influence our inner self and our philosophical understanding of the natural, the artificial, and the definition of humanity.

In each section, we not only discuss the ethical implications and their societal impacts, but also explore potential approaches that could be considered when addressing these issues. We aim to provide researchers, policy makers, and medical workers a guide in developing and applying regenerative medicine—ultimately ensuring that society can maximize the benefits from this new technology.

2. Clinical Translation of Regenerative Medicine

2.1.1 Animal Experimentation in Regenerative Medicine Research

Regenerative medicine relies heavily on animals for testing techniques or providing cells or tissues, raising questions about the ethics of using animals and balancing ethical and scientific concerns.[2,3]

In a review by Vermeulen et. al., it is indicated that animals are capable of experiencing pain and suffering, just as any living creature.[4] Subsequently, utilizing animals in scientific research arguably exploits the suffering and livelihoods of testing animals for the sake of human advancement. The lives of humans are placed above the lives of animals, which is a complicated decision to make. On one hand, the human species possess the distinct capability to execute animal experimentation and have the freedom to advance their well-being as much as possible. On the other hand, there must be a point where research efforts should also take into account its harms: the pain and death caused to animal subjects and even risks posed to human subjects. As life and resources on Earth are limited, it is important to respect the space and other species that we share the Earth with. There is limited indication that humans have a unique right to enact harm on others for their own benefit. Even if humans do, it is important to draw the line as to what extent harm done to animals is permissible.

In evaluating the ethical impact of animal experimentation, multiple perspectives can be taken. Notable moral frameworks include Neo-Rawlsian contractarian theory.[5]
Neo-Rawlsian contractarian theory encourages decisions to be made without the consideration of perspectives such as wealth, status, or even species. In this case, Neo-Rawlsian contractarian theory encourages discussions surrounding animal experimentation to overlook assumed distinctions between human and non-human species. Animal experimentation assumes that humans are somehow inherently superior to non-human animals. At the very least, animal consent is not valued as highly as informed consent from humans. Animal subjects are freely experimented on despite not having the means to achieve informed consent, a crucial ethical rule guiding research on humans. From the baseline that all beings are equal, it is arguable that humans should be tested on just as extensively or, more ethically, animal experimentation with a lower regard for animal suffering or without animal consent should not occur. Though animals do not have the capacity to properly communicate consent or express liberty, those principles can still be assumed and upheld.

Experimentation on animals may not even yield a scientifically meaningful result, further complicating the use of animals. Animal models possess genetic and physiological differences, which may limit its relevance to human physiology.[3,4,6] In fact, adjustments to experimental designs for the purpose of minimizing ethical harm, such as standardization

or reduced sample size, would further restrict an animal experiment's generalizability.

Nevertheless, animal experimentation is regarded by institutions as a crucial step to approve clinical translation. Public perception places humans as more valuable. Additionally, some issues targeted by regenerative medicine, such as simply do not have alternatives, relying solely on animal models.[7] For instance, critical factors in tissue-engineered skeletal muscle (TESM) research include how the TESM interacts with surrounding joints or systems. No computer simulation or cell line captures those interactions as effectively as an animal model.

In order to balance ethical and scientific demands within animal research, the typical approach is referred to as the "3 R's": replace, reduce, and refine. Replacing will be explored later. Focusing on reducing and refining requires extensive consideration of experimental design. An important idea to clarify here is the purpose of animal modeling: collecting evidence about the *in vivo* interactions of regenerative medicine technology to ultimately inform and reduce risks for implementation in human patients.[8] Ideally, the translational distance, or the number of assumptions, between the animal model and human application is reduced. In other words, experimental design will hopefully aim to produce results that benefit the process of clinical translation with limited uncertainty. The 3 R's ensure that animal suffering is also taken into account and minimized in this process.

Researchers will have to determine the smallest number of animals that are harmed that can then be used to generate meaningful results. Specifically, small animals, such as rodents, may be used to refine techniques that will be applied to large animal models before risking multiple large animals.[7] Non-invasive imaging, statistical models, and analgesics are all tools that can be used to minimize suffering.[9] These new technologies should be taken into account when designing, executing, and approving research. For instance, once robust success of tools that can replace animal experimentation is established, regulatory bodies, such as the Food and Drug Administration (FDA), can shift away from requiring animal testing and reframe ethical standards. Gradual shifts in policy can push researchers forward in the reduction of animal experimentation and suffering. Review of regenerative medicine research and technologies could ideally focus on the ethical reduction of animal experimentation in conjunction with the efficacy of the treatment, minimizing unnecessary suffering.

It is important to note that the most impactful solutions still require exploring alternative modeling options and assessing their efficacy and safety for humans, particularly compared to current animal models. The goal is to replace animal models and eliminate the need for them in research, while still informing safe human research and treatments.

2.1.2 Replacing Animal Experimentation

With the advancement of in vitro and computational technology, there are multiple potential alternatives to animal modeling for regenerative medicine. This section attempts to explore the ethical and scientific implications of major examples.

Interestingly enough, regenerative medicine provides alternatives to using animal testing for a variety of applications.[4] Particularly, 3D reconstructed tissue models, including for human skin and cornea, are derived from human cells and directly replicate crucial interactions, such as "barrier function, tissue architecture, and cellular differentiation."[6] Similar technologies are induced pluripotent stem cells, notable in their capability to generate patient-specific models and their contribution to research in genetic disease.[6,10] Further research expanded tissue engineered models into organ-like structures, 3D organoids, as well as tissues with microfluidic structures. Taken together,

tissue-engineered technologies offer multiple pathways to represent distinct physiologies and replace animal models.

A novel 3D dynamic model of angiogenesis, fabricated with layered tubular constructs within a flow chamber, even demonstrated efficacy as a platform to evaluate vascularization of constructs. Results with the model aligned with typical *in vitro* methods.[11] While this is just one example, it highlights how current advancements in regenerative medicine and tissue engineering are going in the direction of accurately modeling human physiology.

Before embracing tissue engineering, though, its accuracy in representing physiological issues and, ultimately, its benefit to clinical translation would need to be further evaluated. Additionally, multiple regenerative medicine techniques rely on human-derived stem cells, prompting questions about consent and ownership. In receiving donations from patients, it is crucial for researchers to be transparent about the specific goals, potential risks, and benefits of stem cell use. Patients rely on a sense of potential benefit and altruism when donating stem cells. Moreover, concerns about privacy or a sense of a part of one's body being immortalized or used can conflict with personal preference.[12] Regenerative medicine research must rely on a robust system of informed consent when using patient-specific biological materials.

Another limitation of tissue-engineered models, particularly in the context of replacing animal models, is their inability to represent systemic interactions. This is where regenerative research often relied on the use of animals to assess integration.

A potential tool to close that gap and reduce dependence on animal models is computational modeling. Artificial intelligence exhibits the capability to handle complex datasets and forecast disease progression.[13] In the case of regenerative medicine modeling, given relevant data from existing animal models, or even patients, deep learning models could potentially be refined to represent systemic interactions. Rather than risking harm to animals, a detailed computer model may suffice. Furthermore, once animal data is adequately detailed, future reliance on animals can be mitigated.

Similarly to the tissue-engineered options, though, computer modeling's efficacy in benefiting clinical translation and reducing risk to humans will need to be further assessed. While specific applications to tissue-engineered heart valves and vascular grafts have found success and approval, relevance can vary case by case and depend on an extensive amount of detailed data.[14] Animal models can reveal nuances in *in vivo* physiology that cannot always be artificially accounted for.[15] Moreover, the use of animal and human patient data within the training of deep learning models will have to be carefully considered. Ethical standards, such as the 3 R's or informed consent, must continue to be upheld. Importantly, regulatory bodies, such as the Institutional Review Board (IRB), must establish a standardized set of ethical guidelines for researchers to follow regarding new technologies. For instance, a computational model must be proven to perform at a specific accuracy or generalizability before being approved for research to ensure its results can confidently be translated to human trials. Moreover, detailed expectations of how animal or human data is handled by computational models should be asserted by review boards and regulatory bodies to maintain informed consent.

Promisingly, a combined approach, incorporating animal models, tissue engineered models, and computational models together, exists as a pathway for shifting away from animal models and navigating the ethical and scientific roles alternative modeling can fulfill and replace in relation to animal models.[15,16]

2.2 In Vivo Human Trials

2.2.1 Informed Consent

The most striking issue in executing in vivo human trials ethically is the lack of broader understanding of regenerative medicine. In a 2023 survey conducted by Padilla et. al., approximately 5,000 US individuals were asked about their familiarity with xenotransplantation.[17] Only 18.3% demonstrated familiarity, highlighting a stark lack of understanding of existing regenerative medicine technologies. Moreover, this survey also reveals important factors in how the public perceives regenerative medicine. When asked about using pig organs to help themselves, discomfort was 41.7%, higher than the 30.8% reported discomfort when asked about using pig organs to help those in need of a transplant generally. Interestingly, "thirty-eight percent of respondents felt more comfortable" with using pig organs once it was revealed that the organs undergo a gene-editing process to increase compatibility.[17] The initial lack of understanding and notable changes in comfort level with added details emphasizes the importance of being properly transparent about scientific information.

A similar gap in scientific understanding is also illustrated in a study comparing the attitudes of members of the public with experts in the field regarding regenerative medicine.[18] Specifically, members of the public were relatively more optimistic about the advancement and benefit of current technology, with 25.9% of experts and a lower 9.0% of members of the public expecting "not much change" or progress in the field within the next 10 years. Naturally, there is a divide between members of the public and experts in the field in scientific understanding of regenerative medicine. When conducting human trials, it is crucial to reduce that divide to appropriately align patient expectations.

Clinical trials inherently carry uncertainty for the outcomes. In order to ensure a clinical trial is conducted ethically, patient rights should be one of the first, if not the first, priorities. Moreover, patient consent and welfare relies on an understanding of the trial—its risks and benefits—and informed consent. Risks include immuno-rejection, lack of vascularization, or unknown long-term effects depending on the specific technology.[2] It is also important to note that, at some point, risks have to be taken. Scientific progress cannot occur without exploring the unknown.[19] Benefit can take on multiple forms of value, such as improving patient and disease health outcomes or advancing regenerative research more broadly. It is critical that the patient thoroughly understands these aspects and provides valid, informed consent.

However, regenerative medicine applications raise complicated challenges to informed consent. For instance, regenerative medicine such as tissue-engineered skin scaffolds are highly integrative, making its implementation relatively permanent. This restricts a patient's ability to voluntarily withdraw from a trial involving tissue-engineered skin scaffolds.[20] In transplantation, similarly, severely injured or trauma patients may not always be able to communicate consent. The decision then heavily relies on the knowledge, integrity, and skills of the researcher or physician as well as regulatory standards. In short, regenerative medicine presents a unique ethical challenge and requires a detailed understanding of its implications to make a morally sound decision.

Importantly, regulatory standards by ethical review boards and institutions, such as the FDA, IRB, or National Institute of Health (NIH) where clinical trials take place should not compromise on informed consent and hold researchers accountable. For instance, it should be required that the patient or volunteer signs off

on a list of details about the treatment provided by researchers and review boards, such as its performance in pre-clinical testing, the potential impacts on the human body, pathways in which the patient or volunteer can opt on before or during the trial, and the reason the regenerative medicine technology is being explored with the patient or volunteer.

Ethical guardrails are important to specify for both review boards and researchers. Approaches to address public knowledge may also include regenerative medicine consult services or even artistic works to educate the public, which will be discussed later.[2,21,22]

2.2.2 Experimental Design

Similar to animal trials, human trials should be considered and designed to minimize risk and maximize benefit.

Both ethically and scientifically, a trial should only be conducted when the risks to the patients are reasonable, especially in relation to potential benefits. For instance, a patient should not risk losing a limb if a specific treatment is poorly understood or is only anticipated to replace necrotic tissue within a small area. In regenerative medicine research, the value of an experiment's results can arise from either direct benefit to a patient's livelihood or an indirect benefit to the broader field's understanding of a specific treatment. To clarify the latter, a diabetic patient, for instance, may participate in a trial of an artificial skin graft on a chronic wound - a regenerative medicine technology - without experiencing significant health benefits. While unfortunate, that case may be subject to analysis, then improve later artificial skin graft treatments. However, this is where challenges may arise. In-human trials may first be conducted on healthy volunteers to assess safety.[23] Within the context of regenerative medicine, though, healthy volunteers may not experience much benefit but still have to experience notable alterations and confront risks.

The same issue applies to participant selection more broadly. Should human trials target patients with issues that are anticipated to directly benefit from the technology? While balancing out the idea of a risk-benefit ratio and potentially maximizing health improvements, targeted participant selection may overlook particular groups, limit generalizability, and skew important results, suggesting that a more diverse pool of participants, including healthy volunteers, is more scientifically sound. Moreover, taking into account the broader value of informing regenerative medicine research described previously may make including healthy volunteers more acceptable. However, that impact is less tangible for volunteers, especially when faced with a potentially drastic or risky procedure, potentially limiting volunteer participation.

Ethical discussions surrounding experimental human trials are also complicated by the idea of personalized tissue-engineered products. While they refine the technology and reduce uncertainty for an individual patient, personalized tissues utilize costly research resources for results with limited generalizability, minimizing broader benefit.[24]

Ethical demands for maximized benefit and minimized risk can come into conflict with scientific demands for extensive and generalizable results and logistical demands for reduced costs. Researchers must strike a balance between personalizing treatments, targeting patients rather than healthy volunteers, and producing generalizable results.

It is arguable that researchers should define a specific group of individuals that their technology is targeted for early on, allowing a clearer translation from benchside to bedside application. For instance, an artificial skin technology may be targeted for chronic wounds in diabetic patients. Researchers will develop

artificial tissue with hyperglycemic environments present in diabetic patients' bodies in mind, optimizing the use of the technology to address a particular issue and defining the scope it can be applied to in clinical trials. Applying the artificial tissue to healthy volunteers that experience normoglycemic environments in their bodies may present unaccounted for risks and may limit their participation in the clinical trial. Delineating a target group for treatment may also allow for additional parameters to provide the basis for pre-clinical research and illuminate a more detailed evaluation of risks and benefits prior to clinical application. Using information on human participants in clinical trials as the basis for research may be enforced by requiring approval of documentation detailing the research protocol and found risks and benefits from institutions such as the FDA and IRB.

Established ethical and regulatory institutions should maintain their role of protecting ethics in the face of rapid technological advancement.

2.3 Clinical Implementation

Despite extensive research and advancements in regenerative medicine, very few successfully cross over into clinical adoption. This translational gap ultimately speaks to a discrepancy in perspectives and limitations due to commercialization.

Preclinical experimentation of regenerative medicine is often driven by elucidating relevant biomechanisms and refining the technology accordingly. Furthermore, ethical discussions are built on the framework that the developed technology should maximize benefit and minimize harm for humanity. Clinical adoption, however, operates entirely differently. While the technology may still be grounded in its goal of providing benefit to humanity, in clinical settings, its success is directly dependent on its compatibility with existing institutions. Ethical concerns come into conflict with logistical challenges.

To clarify, in early research stages, funding is based on the promise of an idea; governments and investors provide money in the hopes that research efforts will mature into something beneficial and ultimately profitable.[25] As a result, experimentation is encouraged to prioritize meaningful and significant results. Already, research can stall there by being unable to prove anything significant and, therefore, lose funding. Particularly for regenerative medicine, financial costs are high, with a clinical-grade iPSC-derived product costing up to $800,000.[24]

Beyond research, though, a technology, such as an iPSC-derived product, must be easily produced and sold to be viable within clinical implementation. Multiple factors not always considered in experimental stages then significantly limit a technology's application, such as regulatory approval, production cost, and efficiency of physician or patient use.[26] In a high-pressure environment like medicine, individuals may be less willing to commit to a high cost or complicated process for the sake of long-term research advancements, which is a more abstract benefit to humanity compared to improving their own livelihoods. Cultural or religious preferences can even oppose technological advancements.[27,28] On top of needing to maximize a perceived benefit and minimize harm, regenerative medicine is also forced to navigate logistical challenges.

A notable concern that arises is whether the demand for regenerative medicine solutions can be met.[29] Regulatory and financial barriers may stand in the way of healthcare access rather than protect individuals from harm. A patient's autonomy to access a tool to benefit well-being should potentially be made more seamless. In general, once regenerative medicine is put into practice, ethical prioritization of human benefit becomes complicated. A "successful" technology can mean one that is scientifically

relevant, ethically beneficial, or practically profitable.

One arguable perspective is that the demands of clinical adoption foster and motivate necessary innovation. As described above, in the research process, it may be more favorable to consider aspects of applications earlier on and optimize the use of resources. Alternatively, rather than trying to balance the roles of consumers or test subjects, members of the public can seek to become "scientific citizens."[29,30] Scientific citizens are empowered beyond simply being allowed informed consent to take on the responsibility of actively engaging in science and shaping research. Science then shifts into a collective social institution where researchers more directly take into account the need to benefit the public rather than being bound by financial interests.

While scientific citizenship is a relatively vague and novel concept, implementation could start at a lower level, taking inspiration from political citizenship and with a branch of the IRB or the NIH overseeing. Patients across clinics and hospitals may receive brief surveys on their regular visits including questions on current and ongoing research at local labs and universities. Here, individuals can learn more about research, present concerns, and propose ideas to researchers, similar to a poll or referendum.

Taken together, research and clinical applications can more easily possess transparency between the public and experts, prioritize public benefit, and incorporate public engagement. Research directions could more easily be dictated by scientific and ethical discussion rather than fiscal interests.

3. Commercial and Regulatory Practices

3.1 Legal Categorization

Traditionally, regenerative medicine could fall into the three following categories: those that are meant to permeate the body with limited and reversible effects (e.g. drugs), those that are forms of corporeal transplantation utilizing naturally-derived blood or tissue (e.g. blood transfusion, organ transplantation), and those that involve insertion of mechanisms or devices (e.g. prosthetic limbs).[31]

However, new kinds of regeneration medicine approaches have emerged in the recent decades that extend beyond the traditional boundaries. They often involve both human or animal tissue in combination with engineered materials. These inserted objects often actively interact with the patient's body to promote cell migration and wound healing, rather than passively existing within the patient's body. For example, certain nanoparticles or chemicals can be incorporated in the scaffold to actively regulate the biological signal transduction pathways.[32] Thus, these regenerative medicines, scaffolds for instance, fall neither into the category of drugs nor medical devices—leaving its legal categorization ambiguous.

Pharmaceutical corporations have taken advantage of this legal categorization failure to evade stricter review or speed up clinical translations. In both the US and the European Union (EU), classifying modern regenerative medicine under the category of prosthetics allows corporations to bypass the stricter review process for drugs, despite regenerative medicine also involving diffusion.[31] The inadequacy of current labels to properly incorporate novel regenerative medicine technologies leads to potential under-regulation. For example, the Celution system produced by Cytori, a US-based company, is an automated system that obtains regenerative cells by processing the patient's adipose tissue. Within the EU, Celution is regulated as a medical device, rather than going through the stricter ATMP framework that would limit the product's profitability. In the US, the device has FDA clearance for use in clinical trials, similarly evading the FDA's more

stringent Cellular and Gene Therapy Pathway regulations.[33]

Addressing this issue is not simply reclassifying or enforcing stricter regulations, however.[34] The principle guiding the EU and US legal system is the substantialist notion of identity, separating technologies into fixed concepts of passive diffusion within the body or physical alterations upon the body.[31,34] This out-of-date assumption led to the legislature's inability to grasp and effectively regulate the dynamic and fluid nature of modern regenerative medicine and their symbiotic relationship with the human body. Existing legal oversights and their potential risk for harm demand a reconsideration of how novel technologies are discussed with respect to a product's relationship with the human body.

3.2 Commercial Narratives

Technology identities are defined as a "narrative or discursive presence of the technology that delineates a particular set of attributed characteristics and performative expectancies as representative of the technology's distinctiveness and value."[35] In other words, institutional and sociopolitical environments often label emerging technologies with "identities", influencing public perception and decision-making. Although these technology identities are not inherently harmful, corporate pursuit of profit can exploit the fluidity of technology identities to spread misinformation and manipulate public perceptions of new treatment options.

For example, the bioengineered trachea is a product aimed at replacing damaged trachea segments, developed in the US by the Inspired Consortium.[33] It received a lot of media coverage—being framed as "addressing a severely unmet clinical need," "lifesaving," "breakthrough," and "pioneering". These technology identities are the results of testimonials given by regulators, supposedly highly reputable surgeons, and the Cell and Gene Therapy Catapult (CGTC). However, this product has become associated with the scandal involving surgeon Paolo Macchiarini who was found to severely misrepresent poor clinical outcomes and is now held under intense scrutiny, contrary to its advertised identity.

3.3 Guiding Public Perception

To counteract biased narratives and encourage the public to adopt a more reasonable and scientific understanding of these emerging regenerative medicines, several methods have been explored. Among these, the utilization of art and consulting services stand out as particularly innovative and promising.

To effectively engage the public, bioethicists can utilize the unique capacity of art to stimulate emotional responses and reshape values. Art, as a sensitive and subtle language, can help raise public awareness of specific ethical issues and encourage people to connect bioethics with their everyday lives. In March 2016 in Montreal, Canada, a non-profit art space Art + Bioéthique took place.[21] The project involved combining artworks, essays and cultural and scientific mediation activities based on the work of six pairs of young bioethics researchers and emerging artists. Bioethicist Victoria Doudenkova reflected in an interview, "Bioethics is a unique field that touches the sensitive string in people. It belongs to everyone, not only to science. Art allows one precisely to seek these sensitive aspects that scientific articles cannot express."[21] This highlights that artworks and bioethics can complement each other in an accessible and meaningful way. However, the extent of the art's influence or quality depends on the relationship forged between researcher and artist, which may differ by case. Nevertheless, this example demonstrates the potential efficacy of employing an interdisciplinary approach to educate the public on bioethics and promote reflection.

In addition to the subtle influences of art, more direct and trustworthy sources of information can also be established. In recent years, with patients increasingly interested in potential regenerative therapies, many sought health information online, which proved to be misleading. Incomplete or dramatized information can influence patients' capacity for reasonable decision making. To address this, Mayo Clinic created a Regenerative Medicine Consult Service (RMCS) in 2011 to help patients navigate the vast amount of information and choose appropriate clinical treatments.[36] Results show that, after the consultation, among patients who were actively pursuing an unproven stem cell intervention (12.9%; n = 65), about a third (32.3%; n = 21) still sought to pursue an unproven intervention toward the end of the consult, experiencing no change in perspectives. However, 16.9% (n = 11) of patients said they would not pursue an unproven intervention at the end of the consult, altering their decision based on details discussed in the consult. The demonstrated impact of proper information provided through consults on decision-making establishes that consult services on regenerative medicine have the potential to reduce the negative effect of the spread of misinformation and encourage the public to adopt a more reasonable perspective on clinical treatments.

4. Societal Impacts of Accessibility and Inaccessibility

4.1. Moral Degradation

Among ethical discussions concerning regenerative medicine, an interesting question is how its accessibility may influence the way individuals regard the value of their physical body.[37,38] Once technologies that repair damaged or diseased cells, tissues, and organs become more widely available, a potential concern is that regenerative medicine may risk contributing to a wider perception that cells, tissues, and organs are "replaceable." For example, activities such as smoking and drinking could be seen as having lower stakes if it is ubiquitously known that cells or organs can be replaced.

Regarding current perception of regenerative medicine, in a survey conducted among patients with Parkinson's Disease, 47.1% of responders "expected regenerative medicine to become the standard therapy within several years," and 60.4% of responders associated regenerative medicine with an improved quality of life.[18] Moreover, cell-based therapies were strongly preferred over treatment options, such as deep-brain stimulation and levodopa-carbidopa intestinal gel. It is important to note that the survey had 102 respondents, potentially limiting its generalizability. Nevertheless, the survey indicates that regenerative medicine is perceived as an attractive and beneficial therapeutic option.

Similarly, in a survey conducted among 5008 participants in the US, the most frequent response (29.2%) for the main point of concern was "lack of current evidence with success and outcome" data.[17] Other primary points of concern include fear of rejection (28.5%) and risk of xenozoonosis (13%). Once regenerative medicine reaches more accessible and safe solutions, such as those being explored by 3D bioprinting or decellularization, a large portion of the concerned population may become increasingly open or reliant on regenerative medicine.

The human body has historically been perceived as fragile and irreplaceable, often shaping risk-averse behaviors. For instance, during the COVID-19 pandemic, risk-averse tendencies and actions regarding health were correlated with increasing age as found in a survey of 130 Italian individuals.[39] In other words, older individuals, with an understanding that their bodies are more vulnerable to severe sickness, took more precautions to protect their health. Interestingly, those with risk-loving

tendencies were less likely to protect their own health. These results highlight how decisions made about personal or public health are influenced by perceived risk.

Regenerative medicine, as a mechanism to improve quality of life or undo damage, fundamentally changes the reality of the human body being irreplaceable and influences lifestyle decisions. Beyond personal decisions, regenerative medicine may even impact how individuals value the physical health of others.

Addressing moral degradation is complex. One's first instinct may be to restrict access to regenerative medicine products, on the basis that moral degradation may only occur on a wide or extreme level if regenerative medicine technology is available in a wide or extreme manner. However, drawing the line on who should or shouldn't receive access to regenerative medicine is difficult.

It may have to be on a case-by-case basis, extending from the target groups of patients defined by researchers discussed in section 2.2.2. Regenerative medicine technology may have to be distributed almost like a prescription medicine, with only patients that meet certain criteria being eligible to receive specific regenerative medicine treatments. The ethicality and efficacy of restricting access is questionable, though, and the potential risk of moral degradation due to regenerative medicine remains an important question to consider.

4.2 Medical Injustice and Social Stratification

An important moral discussion regarding regenerative medicine's accessibility relates to its role as a therapeutic treatment.

Many therapeutic treatments or medical technologies demonstrate benefit only to a specific subgroup of people, particularly groups of higher socioeconomic status.[4] Even with developments in telehealth, marginalized groups continue to suffer. Particularly, low-income, female, and Black patients experience a gap in healthcare access.[40] Moreover, this occurs on both a smaller scale—within a city or state—and a larger scale—across the nation or even the world. Individuals in rural areas or low-income or middle-income countries experience poorer access to healthcare and poorer health outcomes. It is an unfortunate reality that access to healthcare is notably impacted by the demographics of the area one lives in. Moreover, disparities in healthcare access also further exacerbate poverty and socioeconomic stratification, further marginalizing disadvantaged groups of people.[41]

Aside from therapeutic options, medical research also experiences bias. The unfortunate reality is that regenerative medicine is expensive and most accessible to those who can afford it.[23] Tissue-engineered skin scaffolds cost thousands of dollars.[20,42] Similarly, stem cell technology relies on specialized infrastructure, likely limiting access to those in the proximity of an urban research facility.[43] Moreover, stem cell research relies on data from individual biology, meaning those who most easily access stem cell technologies will also be the ones most effectively represented and supported in ongoing research.

In short, regenerative medicine may further widen the gap between socioeconomic status in terms of quality of life and health outcomes.

Minimizing discrepancies between socioeconomic groups in medical research and treatment may go hand-in-hand with the scientific citizenship idea mentioned in section 2.3, where surveys about regenerative medicine are delivered in a variety of clinics or hospitals or even conducted online. Furthermore, local hospitals may receive funding from the NIH or governmental bodies to enable a variety of patients to participate in trials or research.

4.3 The Role of Regenerative Medicine

The current field of regenerative medicine focuses on restoring normal function for damaged cells, tissues, or organs. Core techniques include stem cell therapy, tissue engineering, and transplantation. Those significantly impacted by developments often include cancer patients, diabetic individuals, severe burn or injury victims, or other patients afflicted by genetic conditions, creating a unique need.[44,45,46]

It is important that the patients most impacted by regenerative medicine suffer from conditions that are otherwise very difficult to treat. The right to live without suffering is a primary goal of regenerative medicine and medicine overall. While some areas, such as weight loss, may have cultural complexities, regenerative medicine is grounded in directly improving health and quality of life.

It addresses an interesting debate within medical treatment regarding prevention versus treatment. Certain health conditions are unavoidable, and there still exists a crucial need for therapeutic options.

Moreover, it allows those working in the field to have a clear idea of a specific benefit that can and should be explored. Those with certain conditions should not have a more difficult time accessing economic or social opportunities. Even those without certain conditions who interact with regenerative medicine need an understanding of the advantages and disadvantages it has, which is what makes development and discussion in regenerative medicine so important.

5. Impacts on Human Identity

5.1 Natural vs. Artificial

It has long been taken for granted by the public that there is a clear boundary between the natural and the artificial. Yet, discussion surrounding the distinction is crucial in regenerative medicine, where advancement, especially in tissue engineering, inherently blurs that boundary.[28] Particularly, the engineered scaffold is artificial, physically alien to the human body. However, it can also be interpreted as natural, since the patient's own cells will migrate into the scaffold in the process of tissue regeneration. The scaffold may even be degraded and replaced by the patient's own extracellular matrix in later stages. Therefore, the regenerated tissue is made up of both artificial and natural parts. As a result, unlike previous medical devices such as bionic technologies, regenerative medicine redefines how we view the concept of being natural. The subtle relationship between the patients and their regenerated body parts need to be taken into consideration by researchers and medical staff, as a sense of confusion or estrangement towards their own body can negatively impact rehabilitation outcomes.[23]

5.2 Perception of Self

The advancement of regenerative medicine also brings what defines us as human beings into question, encouraging further discussions on human embodiment. Historically, philosophers have long held the belief that both or either the intact biological body and the mind, or our thoughts and moral values, play a significant role in the construction of self. However, philosophers' opinions on mind-body duality and the continuity of self differ, offering varying perspectives with which to consider the concept of self and how regenerative medicine may impact it.

Plato believed that the soul is distinct from the biological human body and exists independently.[47] The persistence of self relies on the soul, rather than the temporary body. Therefore, alterations to the biological body do not affect our perception of self. René Descartes shared similar ideas, believing that the body and mind are distinct, while the consciousness defines self.[48] In contrast, Aristotle believed that

the mind and the body are deeply connected, where the internal source of motion ultimately stems from the heart.[48] According to Aristotle, the personal identity is only consistent when the organismal body is consistent, suggesting that employing regenerative medicine alters self.

Following a more contemporary or fluid line of thinking, John Locke believed that personal identity depends on psychological consistency, or the consistency of one's consciousness.[49] This suggests that tissue regeneration does not alter one's perception of self. However, when discussing the wider term regenerative medicine which involves the development of artificial organs, the brain in particular, will change one's psychological processes, and the concept of oneself. In a slightly distinct manner, David Hume and Derek Parfit believed that perception of self is not stationary, but develops and changes as our perspectives and experiences change.[49]

5.3 Beyond Human

As regenerative medicine develops, we may come to a stage where the technology enables individuals to become more than humans, outgrowing its current role in treating humans with disease. Humans inherently have the desire to improve themselves to the extent that exceeds normal human capabilities. Especially when considering the anti-aging potentials of regenerative medicine, we may enter a "transhuman" phase and the concept of "human being" itself becomes out of date.[50]

When applying regenerative medicine, it is important to consider the personal beliefs of the patients and potentially even establish broader societal standards on an acceptable degree of "self-alteration."

6. Conclusion

While scientific development often comes with uncertainty, encouraging extensive discussions and engagement as well as establishing feasible systems to move forward allows progress to continue in an ethical and meaningful way. In this review, the field of regenerative medicine was examined from an ethical lens.

With respect to clinical translation, specific areas of ethical concern include animal experimentation, in vivo human trials, and clinical implementation. Minimizing harm and maximizing benefit quickly becomes complicated when refining regenerative medicine technologies. Potential approaches to address these ambiguities follow the principles of "replace, reduce, and refine" when conducting animal experimentation, replacing animal experimentation with 3D constructed models or computational modeling, ensuring transparent communication between researchers and patients, and guiding the public into becoming "scientific citizens,"

In terms of commercial and regulatory practices, the ethicality of regenerative medicine applications is placed at risk due to flawed legal categorizations and misleading commercial narratives. Patients may develop a skewed understanding of therapies, violating their right to informed consent or manipulating their decision making unfairly. To address these issues, regulatory bodies can reconsider the relationship between the human body and pharmaceutical products; providers of regenerative medicine can establish consulting services or even incorporate artworks to educate the public on bioethics and encourage reflection.

Ethical concerns also arise when considering the societal impacts of regenerative medicine's accessibility or distribution. Societies may experience moral degradation, medical injustice, and stratification. It is critical to specify the intended role of regenerative medicine as well as discuss and understand those unintended. Important ideals to maintain include guiding the public to form a new understanding of physical health, increasing the accessibility of

regenerative medicine to disadvantaged groups, and fostering continuous discussion on the development of regenerative medicine.

Lastly, regenerative medicine challenges abstract conceptualizations of human identity, shifting definitions and discussions. Ethical concerns include the blurring of boundaries between the natural and the artificial, the struggle of forming a continuous self, and the pursuit of the "transhuman" phase. Potential solutions include fostering continuous discussions on the development of regenerative medicine.

Overall, scientific research must align with public perception, perform experiments with a clear purpose or goal, and be grounded in relevant transparency and informed consent. Ethical discussions are crucial to ensure scientific research provides the maximum possible benefit to humanity.

7. References

[1] Mason, Chris and Peter Dunnill. "A Brief Definition of Regenerative Medicine." *Regenerative Medicine*, vol. 3, no. 1, 2007, pp. 1–5, https://doi.org/10.2217/17460751.3.1.1.

[2] Patuzzo, Sara, et al. "3D Bioprinting Technology: Scientific Aspects and Ethical Issues." *Science and engineering ethics*, vol. 24, no. 2, 2018, pp. 335-348. https://doi.org/10.1007/s11948-017-9918-y

[3] Wakeman, Dustin R, et al. "Large Animal Models Are Critical for Rationally Advancing Regenerative Therapies." *Regenerative Medicine*, vol. 1, no. 4, 2006, pp. 405–13, https://doi.org/10.2217/17460751.1.4.405.

[4] Vermeulen, Niki, et al. "3D bioprint me: A socioethical view of bioprinting human organs and tissues." *J Med Ethics*, vol. 43, no. 9, 2017, pp. 618–624, https:// doi.org/ 10.1136/medethics-2015-103347

[5] Liguori, Gabriel, et al. "Ethical Issues in the Use of Animal Models for Tissue Engineering: Reflections on Legal Aspects, Moral Theory, Three Rs Strategies, and Harm–Benefit Analysis." *Tissue Engineering Part C: Methods*, vol. 32, no. 12, 2017, pp. 850-862, https:// doi.org/10.1089/ten.tec.2017.0189

[6] Lee, Chang-Jin, et al. "Advanced Animal Replacement Testing Strategies Using Stem Cell and Organoids." *International Journal of Stem Cells*, Korean Society for Stem Cell Research, 2025, https://doi.org/10.15283/ijsc24118.

[7] Baker, Hannah, et al. "Ethical considerations in tissue engineering research: Case studies in translation." *Methods (San Diego, Calif.)* vol. 99, 2016, pp. 135-44. https://doi.org/10.1016/j.ymeth.2015.08.010

[8] Karami, Peyman, et al. "A Guide to Preclinical Evaluation of Hydrogel-Based Devices for Treatment of Cartilage Lesions." *Acta Biomaterialia,* vol. 158, 2023, pp. 12-31, https://doi.org/10.1016/j.actbio.2023.01.015.

[9] Anup, Amritha, et al. "Embracing Ethical Research: Implementing the 3R Principles into Fracture Healing Research for Sustainable Scientific Progress." *Journal of Orthopaedic Research*, vol. 42, no. 3, 2023, pp. 568–77, https://doi.org/10.1002/jor.25741.

[10] Park, Guiyoung, et al. "Replacing Animal Testing with Stem Cell-Organoids : Advantages and Limitations." *Stem Cell Reviews and Reports*, vol. 20, no. 6, Springer Nature, 2024, https://doi.org/10.1007/s12015-024-10723-5.

[11] Dikici, Serkan, et al. "Assessment of the Angiogenic Potential of 2-Deoxy-D-Ribose Using a Novel in Vitro 3D Dynamic Model in Comparison With Established in Vitro Assays." *Frontiers in Bioengineering and Biotechnology*, vol. 7, Frontiers Media SA, 2020, https://doi.org/10.3389/fbioe.2019.00451.

[12] Dasgupta, Ishan, et al. "Patients' Attitudes toward the Donation of Biological Materials for the Derivation of Induced Pluripotent Stem Cells." *Cell Stem Cell*, vol. 14, no. 1, 2014, pp. 9–12, https://doi.org/10.1016/j.stem.2013.12.006.

[13] Nosrati, Hamed, and Masoud Nosrati. "Artificial Intelligence in Regenerative Medicine: Applications and Implications." *Biomimetics (Basel, Switzerland)* vol. 8, 5 442. 20 Sep. 2023, https://doi.org/10.3390/biomimetics8050442

[14] Loerakker, Sandra, and Jay D. Humphrey. "Computer Model-Driven Design in Cardiovascular Regenerative Medicine." *Annals of Biomedical Engineering*, vol. 51, no. 1, Aug. 2022, pp. 45–57, https://doi.org/10.1007/s10439-022-03037-5.
[15] Mahalmani, Vidya, et al. "Do alternatives to animal experimentation replace preclinical research?." *Indian Journal of Pharmacology*, vol. 55, no. 2, Mar. 2023, pp. 71–75, https://doi.org/10.4103/ijp.ijp_223_23.
[16] Mittal, Rahul, et al. "Exploring the Potential of Computer Simulation Models in Drug Testing and Biomedical Research: A Systematic Review." *Frontiers in Pharmacology*, vol. 16, Sept. 2025, https://doi.org/10.3389/fphar.2025.1644907.
[17] Padilla, Luz A, et al. "Public Attitudes to Xenotransplantation: A National Survey in the United States." *American Journal of Transplantation*, vol. 24, no. 11, Elsevier BV, July 2024, https://doi.org/10.1016/j.ajt.2024.07.018.
[18] Shineha, Ryuma, et al. "A Comparative Analysis of Attitudes on Communication Toward Stem Cell Research and Regenerative Medicine Between the Public and the Scientific Community." *Stem Cells Translational Medicine*, vol. 7, no. 2, Jan. 2018, pp. 251–57, https://doi.org/10.1002/sctm.17-0184.
[19] Hermerén, Göran. "The ethics of regenerative medicine". *Biologia Futura*, vol. 72, 2021, pp.113–118. https://doi.org/10.1007/s42977-021-00075-3
[20] Hunckler, Michael, and Aaron Levine. "Navigating Ethical Challenges in the Development and Translation of Biomaterials Research." *Frontiers in Bioengineering and Biotechnology*, vol. 10, Sept. 2022, https://doi.org/10.3389/fbioe.2022.949280.
[21] Couture, Vincent, et al. "Merging Arts and Bioethics: An Interdisciplinary Experiment in Cultural and Scientific Mediation." *Bioethics*, vol. 31, no. 8, Wiley, Sept. 2017, pp. 616–30, https://doi.org/10.1111/bioe.12391.
[22] Terzic, Andrew, et al. "Regenerative Medicine Build-Out." *Stem Cells Translational Medicine*, vol. 4, no. 12, Dec. 2015, pp. 1373–79, https://doi.org/10.5966/sctm.2015-0275.
[23] Van Daal, Manon, et al. "Personalized 3D Printed Scaffolds: The Ethical Aspects." *New Biotechnology*, vol. 78, no. 78, Dec. 2023, pp. 116–22, https://doi.org/10.1016/j.nbt.2023.10.006.
[24] Lin, Haishuang, et al. "Integrated Generation of Induced Pluripotent Stem Cells in a Low-Cost Device." *Biomaterials*, vol. 189, Jan. 2019, pp. 23–36, https://doi.org/10.1016/j.biomaterials.2018.10.027.
[25] Velcani, Frida, et al. "Tissue and Organ Engineering: Investor Sentiments and Analysis." *Regenerative Therapy*, vol. 31, Mar. 2026, p. 101066, https://doi.org/10.1016/j.reth.2026.101066.
[26] O'Donnell, Benjamen T., et al. "Beyond the Present Constraints That Prevent a Wide Spread of Tissue Engineering and Regenerative Medicine Approaches." *Frontiers in Bioengineering and Biotechnology*, vol. 7, May 2019, https://doi.org/10.3389/fbioe.2019.00095.
[27] de Kanter, Anne-Floor J., et al. "The Ethical Implications of Tissue Engineering for Regenerative Purposes: A Systematic Review." *Tissue Engineering Part B: Reviews*, vol. 29, no. 2, Oct. 2022, https://doi.org/10.1089/ten.teb.2022.0033.
[28] Vijayavenkataraman, S., et al. "3D Bioprinting – An Ethical, Legal and Social Aspects (ELSA) Framework." *Bioprinting*, vol. 1–2, Elsevier BV, Mar. 2016, pp. 11–21, https://doi.org/10.1016/j.bprint.2016.08.001.
[29] Chan, Sarah. "Research Translation and Emerging Health Technologies: Synthetic Biology and Beyond." *Health Care Analysis*, vol. 26, no. 4, Dec. 2016, pp. 310–25, https://doi.org/10.1007/s10728-016-0334-2.
[30] Woolley, J. Patrick, et al. "Citizen Science or Scientific Citizenship? Disentangling the Uses of Public Engagement Rhetoric in National Research Initiatives." *BMC Medical Ethics*, vol. 17, no. 1, June 2016, https://doi.org/10.1186/s12910-016-0117-1.
[31] Parry, Bronwyn. "The Social Life of 'Scaffolds.'" *Science, Technology, & Human Values*, vol. 43, no. 1, SAGE Publications, Oct.

2017, pp. 95–120, https://doi.org/10.1177/0162243917735179.
[32] Dahiya, Aditi, et al. "3D Printed SiO_2 –Tricalcium Phosphate Scaffolds Loaded with Carvacrol Nanoparticles for Bone Tissue Engineering Application." *Journal of Medicinal Chemistry*, vol. 67, no. 4, Dec. 2023, pp. 2745–57, https://doi.org/10.1021/acs.jmedchem.3c01884.
[33] Gardner, John, et al. "Promissory Identities: Sociotechnical Representations & Innovation in Regenerative Medicine." *Social Science & Medicine*, vol. 174, Feb. 2017, pp. 70–78, https://doi.org/10.1016/j.socscimed.2016.12.018.
[34] Fisch, Michael. "Regenerating Bodies." *Science, Technology, & Human Values*, vol. 43, no. 1, SAGE Publications, Oct. 2017, pp. 121–28, https://doi.org/10.1177/0162243917737364.
[35] Ulucanlar, S., et al. "Technology Identity: The Role of Sociotechnical Representations in the Adoption of Medical Devices." *Social Science & Medicine*, vol. 98, Elsevier BV, Dec. 2013, pp. 95–105, https://doi.org/10.1016/j.socscimed.2013.09.008.
[36] Smith, Cambray, et al. "Challenging Misinformation and Engaging Patients: Characterizing a Regenerative Medicine Consult Service." *Regenerative Medicine*, vol. 15, no. 3, Mar. 2020, pp. 1427–40, https://doi.org/10.2217/rme-2020-0018.
[37] Vijayavenkataraman, Sanjairaj. "A Perspective on Bioprinting Ethics." *Artificial Organs*, vol. 40, no. 11, Nov. 2016, pp. 1033–38, https://doi.org/10.1111/aor.12873.
[38] Jamieson, Colin, et al. "A Review of Recent Advances in 3D Bioprinting With an Eye on Future Regenerative Therapies in Veterinary Medicine." *Frontiers in Veterinary Science*, vol. 7, Feb. 2021, https://doi.org/10.3389/fvets.2020.584193.
[39] Cerami, Chiara, et al. "Risk-Aversion for Negative Health Outcomes May Promote Individual Compliance to Containment Measures in Covid-19 Pandemic." *Frontiers in Psychology*, vol. 12, June 2021, https://doi.org/10.3389/fpsyg.2021.666454.
[40] Saeed, Sy Atezaz, and Ross MacRae Masters. "Disparities in Health Care and the Digital Divide." *Current Psychiatry Reports*, vol. 23, no. 9, July 2021, p. 61, https://doi.org/10.1007/s11920-021-01274-4.
[41] Abdi, Yusuf Hared, et al. "Understanding Global Health Inequality and Inequity: Causes, Consequences, and the Path Toward Justice in Healthcare." *Public Health Challenges*, vol. 4, no. 4, Wiley, Oct. 2025, https://doi.org/10.1002/puh2.70156.
[42] Vig, Komal et al. "Advances in Skin Regeneration Using Tissue Engineering." *International journal of molecular sciences* vol. 18, no. 4, pp. 789. 7 Apr. 2017, https://doi.org/10.3390/ijms18040789
[43] Yap, Kiryu K. "Inequality Issues in Stem Cell Medicine." *Stem Cells Translational Medicine*, vol. 5, no. 2, Dec. 2015, pp. 265–66, https://doi.org/10.5966/sctm.2015-0232.
[44] Arthurs, Jennifer R., et al. "Patients Seeking Stem Cell Therapies—a Prospective Qualitative Analysis from a Regenerative Medicine Consult Service." *Npj Regenerative Medicine*, vol. 7, no. 1, Mar. 2022, https://doi.org/10.1038/s41536-022-00215-w.
[45] Mao, Angelo S., and David J. Mooney. "Regenerative Medicine: Current Therapies and Future Directions." *Proceedings of the National Academy of Sciences*, vol. 112, no. 47, Nov. 2015, pp. 14452–59, https://doi.org/10.1073/pnas.1508520112.
[46] Hussen, Bashdar Mahmud, et al. "Revolutionizing Medicine: Recent Developments and Future Prospects in Stem-Cell Therapy." *International Journal of Surgery*, vol. 110, no. 12, Elsevier BV, Nov. 2024, https://doi.org/10.1097/js9.0000000000002109.
[47] Calef, Scott. "Dualism and mind." *Internet Encyclopedia of Philosophy* 9, 2005.
[48]Dempsey, Liam P. "Thinking-Matter Then and Now: The Evolution of Mind-Body Dualism." *History of Philosophy Quarterly*, vol. 26, no. 1, 2009, pp. 43–61. *JSTOR*, http://www.jstor.org/stable/27745145.

[49] Sedikides, Constantine, et al. "Self-continuity." *Annual review of psychology*. vol, 74, 2023, pp. 333-361. https://doi.org/10.1146/annurev-psych-032420-032236

[50] Moreno, Jonathan, et al. "Managing Cell and Human Identity." *Science*, vol. 356, no. 6334, American Association for the Advancement of Science, Apr. 2017, pp. 139–40, https://doi.org/10.1126/science.aan2763.

Journal of Ethics in Scientific and Technological Innovations

Therapy Versus Enhancement: The Ethics of Emerging Transhumanist Technology

Marie Kouzmine[1]
[1]*The Bear Creek School, Redmond, WA 98053*

Abstract

In recent years, there have been emerging advances in the scope of effectiveness of medicine and technology for human application. Engineered bio-integrated sensors and neural interface polymers, for instance, integrate synthetic products directly into the human body. Yet these novel materials simultaneously blur the line between therapy and transhumanism. In this paper, the implications of integrating materials with the human body are considered. Specifically, ethical implications for the individual and population as a whole, as well as a regulatory paradigm for transhumanist technologies are highlighted.

1. Introduction

As medicine and technology grow more effective at interacting with the systems of the human body, the potential is unlocked for not just restoring function to systems when they break, but improving them beyond what is natural.

This avenue has vast ethical implications and can be predicted to affect both individuals and the human population as a whole, as will be explored in this paper. These concerns mandate that a strategy for handling transhumanism be developed while transhumanist technology is still in its infancy. Such options must take into account both the prevention of ethical concerns and the continuation of human flourishing.

2. Situational Overview

First, a brief overview of the concepts and definitions used in this paper will be presented, as well as the context in which this paper is relevant.

2.1 Therapeutic Medicine Versus Enhancement

The term "therapeutic medicine" is used as an opposite characterization to enhancement. Therapy focuses on restoring the baseline function of the human body in reaction to an abnormal (not self-induced, not typical of the majority of humans) worsening of condition.[1] Hence, in this paper, "baseline function" refers to the original state of the body prior to the abnormal worsening. For instance, taking corticosteroids for allergy relief returns one's breathing to regular levels, just as traditional prosthetic legs restore one's natural walking ability.[2] It is important to note that this paper therefore considers baseline function to vary from person to person: if a person who can deadlift 150 pounds is affected by a muscular-degenerative disease, therapy would ideally restore their body to being once again capable of lifting 150 pounds; if a person who can lift 50 pounds were to undergo the same disease, therapy would ideally restore their ability to lift this baseline of 50 pounds, no more.

In contrast to therapy, "enhancement" means improving one's traits beyond what is necessary for restoration of original health.[3] Enhancement is therefore generally not a reaction to an abnormal downturn in health, but a proactive choice on the part of the subject.

2.2 Definition of Transhumanism

Transhumanism is a philosophy that embraces enhancement for the evolution of the human species. Transhumanists seek to "augment human capabilities" in pursuit of what they believe to be the benevolent cause of improving the human condition, also known as eudaimonia.[4] To do this, they propose the use of technologies such as "genetic engineering, cryonics, artificial intelligence, and nanotechnology."[4]

2.3 Modern Relevance

Due to the currently burgeoning and poorly regulated nature of transhumanist technology and the powerful support it is now receiving, it is relevant and crucial to determine how it should be viewed.

Firstly, transhumanist projects have "support from Silicon Valley entrepreneurs, including Google cofounder Larry Page, Amazon's Jeff Bezos, and Tesla's Elon Musk."[4] With such an influx of investment entrepreneurs, research into transhumanist technologies will only expand. The aforementioned Neuralink brain chips, supported by Elon Musk, are one example of this.

Indeed, as is pointed out by Francis Fukuyama, a leading bioethicist, transhumanism is already implicitly present in much of contemporary advanced research.[6] For instance, androgenic-anabolic steroids (AAS) are synthetic hormones that increase strength by about 5-20% with no physical training required from the subject.[23] Similarly transcranial direct current stimulation (tDCS) devices improve concentration and problem-solving through providing electrical current to the brain.[24]

Yet despite this present relevance, there is a regulatory gap for such technology. The United States Food and Drug Administration (FDA) only regulates devices that are intended to interact with diseases or have medical implications.[24] Another regulatory body, the Consumer Product Safety Commission (CPSC) focuses on handling physical hazards of products.[24] Transhumanist technologies are able to exploit these limitations in order to enter the market without regulation. For instance, tDCS products are publicly available, out of reach of the FDA because they are marketed towards healthy people, and not investigated by the CPSC since they have no physical hazards.[24] Thus, the mental hazards of tDCS have escaped evaluation. This poor regulation for technology that will only continue to grow in prominence is concerning.

As a whole, the emerging and supported growth of transhumanist technologies coupled with faulty regulation prompt widespread deliberation on whether transhumanism is, as some put it, "the world's most dangerous idea."[5]

3. Ethical Concerns of Transhumanism

This section lays out a diverse array of ethical concerns brought on by aspects of transhumanism, ranging from the global to individual scale.

3.1 Societal Stratification

One of the chief concerns about transhumanist technologies is that they will lead to an increase in societal stratification through exacerbating existing divisive factors such as biology and age, ultimately leading to an erosion of equal rights.

3.1.1 Inequality of Access

For one, it can be predicted that transhumanist technologies will be inaccessible to a large portion of the population, namely the citizens of the world's poorest countries.[6] In the modern day, there are people in the world who do not have access to basic technologies, like plumbing, which have already existed for centuries. The World Health Organization and United Nations Children's Fund report that 2.1 billion people still lack access to safe drinking

water, and 3.4 billion lack proper sanitation.[7] Accordingly, it is not reasonable to assume that enhancement technologies, far more advanced and costly than plumbing, would be universally available anytime soon after their creation. In consequence, the wealthy would be able to turn their money into a permanent biological advantage, further physically separating the elites from the common people.

3.1.2 Obsolete Generations and Overpopulation

One of the most commonly pursued goals of transhumanism is extending the human lifespan.[8] Although stopping aging would indeed be revolutionary, it poses multiple ethical concerns.

Leon Kass, who was the chairman of the President's Council on Bioethics, asks the following: "what incentive would there be for the old to make way for the young, if the old slowed down but little and had no reason to think of retiring?"[9] If lifespans were extended, and the perverse effects of aging stopped, some hypothesize that the younger generation would forever be second to the older in knowledge and experience, and would be made effectively obsolete. Moreover, with the vast decrease in deaths from old age, overpopulation would be accelerated and living space could become a premium. Those old enough to have been alive before the population boom might therefore have the additional advantage of property ownership. Thus, eliminating aging risks sharpening the societal divide created by age.

3.1.3 Erosion of Human Rights

If this growth in divisions were to manifest, the castes of human—unenhanced persons—and more human—enhanced persons—could fracture the unified identity of humanity. This in turn would risk eroding universal human rights, which find their foundation in "the belief that we all possess a human essence that dwarfs manifest differences."[6]

3.2 Identity

Transhumanist advances also pose severe potential risks to human identity and personhood due to their introduction of more degrees of freedom for personal modification.

3.2.1 Promotion of Conformity

Fukuyama poses an essential question about transhumanism, stating, "if some move ahead, can anyone afford not to follow?"[6] Since enhancements would improve the user's abilities, living without them would become extremely disadvantageous. Smartphones are one existing example of this principle: their enhancement of communication has become a requirement of living in modern society. People who live without them are excluding themselves from participation in a "normal" life of rapid information access and vast social network, putting them at an extreme disadvantage.[10]

Additionally, transhumanist technologies strive to increase the totality of ways one can alter themselves. The more control people have over their traits, the greater the extent to which homogenization can occur. Consider modern cosmetics: due to technologies like Botox, fillers, and rhinoplasty increasing one's control over their face, modern celebrities "end up looking like slightly different variations of the exact same face."[11] Similarly, the self-customization options transhumanism aims to produce might drastically decrease the diversity of humanity.

3.2.2 Commodified Self

Furthermore, technology that integrates directly with the human body blurs the line between what is human and what is machine. For instance, if memory and cognitive ability, core aspects of consciousness, become something that can be edited over the course of

one's life, it could recenter one's identity from being based in life experiences to being rooted in technical upgrades. One would become a "theorized and fabricated" being.[12] For such a person, self-worth in turn risks becoming tied to optimization, rendering the biological self an outdated baseline.

3.3 Autonomy

There are also concerns about how autonomy would be affected in a transhumanist future.

3.3.1 Direct Cognitive Influence

Because the goal of enhancement is optimization of human qualities, including cognitive qualities, free will may be negatively affected.[8] The key difference between existing cognitive aids and permanent mental enhancement is in their method of use. For instance, artificial intelligence is commonly criticized as a tool that does the thinking for its users, hindering their ability to make autonomous decisions.[13] Yet the user still has the conscious choice of whether they use or do not use the response AI provides them. As such AI is an indirect cognitive aid.

Certain transhumanist technologies might instead aim to directly send signals to one's brain, artificially manipulating neuron signals in a way that contradicts free will (the ability to take actions without outside manipulation).[14] One burgeoning example of this is deep brain stimulation, which uses electrodes implanted within the brain to send electrical impulses to brain regions. A study found that 33% of patients felt an "existential crisis of agency" due to how the procedure operated their body through an external device, making their actions not their own.[15] This predicts a serious concern about one of the avenues transhumanist technology has already begun to take.

3.3.2 Technological Vulnerabilities

Integrated technologies also introduce digital vulnerabilities, providing a way for one's body to be taken over and directed remotely, or for one's personal information to be extracted without their consent or knowledge.[16] This transforms one's body from their one true possession to an exploitable part of a network.

3.4 Human Striving

Finally, the ethical implications of transhumanism on human striving, or pursuit of curiosity and improvement, must be taken into account.

3.4.1 Homogenized Experience

Firstly, the aforementioned increase in societal homogenization will in turn reduce "the likelihood of genuine freedom, individuality, and greatness."[9] One of the sources of human striving comes is the originality that variance in experiences and struggles conveys upon people: it is the differences between people's lives that support unique passions and interests. If everyone converges towards the same traits, living similar lives, originality and consequently striving could be negatively impacted.

3.4.2 Effortless Living

Beyond this, as Kass points out, "biomedical interventions [will] act directly on the human body and mind to bring about their effects on a subject who … plays no role at all."[9] When there is a drug to boost muscle mass, what reason is there to exercise at all? When there is a drug for endless euphoric happiness, why would a person ever pursue their passions? Of course, self-motivated people exist, and not all transhumanist technologies would have such an effect. Still, those transhumanist technologies that provide the benefits of hard work without any effort from the user would simultaneously increase human passivity, which would even further reduce human striving.

3.4.3 The Value of Striving

Throughout history, human striving has been pervasively identified as a requirement for a meaningful life. Thousands of years ago among the ancient Greeks, Aristotelian virtue ethics established that the value in life comes from recognizing, harnessing, and cultivating one's unique human functions.[17] Most recently, Existentialism maintains "existence precedes essence," and Absurdism that life is inherently meaningless, upon which both philosophies argue that meaning must be created through effort.[18,19] Though not all people believe effort creates meaning, such a philosophy is prominent enough in the modern world that the possible reduction of striving through homogenization and effortless living would have a non-trivial impact on the psychological security of the world.

4. Evaluation of Regulation Options

With transhumanism spurring such a variety of ethical concerns and being on the verge of hitting society full force, a practical course of action must be set out for its regulation. This section will explore the commonly suggested options and recommend one for policymakers.

4.1. Popular Regulation

Radical proponents of transhumanism suggest no course of action is necessary, rather that popular regulation will suffice in keeping transhumanist technologies on the right course, but this is refuted by societal and historical evidence. Stephen Lilley, a professor in favor of transhumanism, claims, "if certain applications appear too dangerous or repulsive, consumers will not support them and they will fail in the market" and "free citizens will not opt for technologies that limit self-determination."[20] These statements may be true of a population of ideally logical humans (after all, what person would willingly choose oppression and harm), but people are not ideally logical. Over a billion people across the globe use tobacco, even though doing so causes cancer and heart disease.[21] Moreover, as was evident in the Russian Revolution, it is possible for millions of people to willingly support a movement that ends up taking their freedom away.[22] Ultimately, a definitive response to transhumanism must be considered; leaving transhumanism to popular regulation will not ensure its safe development or the avoidance of the aforementioned ethical concerns.

4.2. Complete Prohibition

On the opposite side, complete prohibition proposes that enhancement technology and medicine be banned entirely from the commercial and research spheres. This strategy successfully circumvents the majority of the ethical issues, although it is important to establish that complete prohibition is not realistically possible. Throughout history, smuggling and illegal production of goods has accompanied any prohibition: for instance, the widespread drug smuggling business operational to this day, or alcohol smuggling during American Prohibition. Inevitably a small portion of the population will still acquire and use transhumanist technologies. Even still, identity, autonomy, and human striving would be protected because the vast majority of people would not acquire such technology. And though it is likely most elites would enhance themselves, the clandestine nature of such activity would reduce the societal stratification it inflicts.

However, complete prohibition fails in its denial of possible increases to human quality of life. For instance, increasing the human lifespan is generally desirable, as many people would want to live longer, or for their loved ones, parents, or grandparents to. Of course, as earlier discussed, extending life spans comes

with a variety of ethical concerns. However, is it really worth stagnating human development permanently in fear of these concerns?

4.3. Preparative Withholding

Preparative withholding, the recommendation of this paper, strikes a compromise between the previous two strategies, directing that transhumanist technology be withheld from the commercial market, but not research, until society is ready to integrate it. It is built on the observation that the ethical issues brought about by transhumanism are not necessarily inherent to transhumanism itself, but to society's unpreparedness for transhumanism.

Take, for example, stratification caused by the lack of living space in a planet populated by near immortal humans. Ultimately, the issue is not that people do not age; it is that Earth has a limited amount of space. If the viability and existence of colonies outside of the planet were ensured before longevity products became publicly available, the issue would not manifest.

In this way, preparative withholding can both address ethical problems of transhumanism while simultaneously allowing the movement towards eudaimonia to continue.

5. Conclusion

Transhumanism—the philosophy of enhancing human capabilities beyond what is natural—will only become more relevant as time goes on and technology and medicine continue to advance. Already, transhumanist technologies are being developed and backed by wealthy investors. As such, the effects of transhumanism must be rigorously considered.

A variety of possible ethical concerns, for both the population as a whole, and for individuals, have been identified in this paper. Current regulatory bodies are insufficient to ensure the safety of such technologies. Moreover, the variety of ethical concerns described indicate popular regulation will not be sufficient to ensure the safe development of transhumanist technologies in the future. Though complete prohibition would prevent the majority of these concerns, this paper proposes that preparative withholding is overall superior due to it achieving a similar level of protection while staying open to the potential of transhumanist technologies to improve human quality of life.

6. References

[1]Kuhse, Helga, and Peter Singer. Bioethics : An Anthology. Blackwell Pub, 2006.

[2]Pellegrino, Edmund. "Biotechnology, Human Enhancement, and the Ends of Medicine." Dignitas, vol. 29, no. 3–4, 2022, pp. 15–17, https://www.cbhd.org/dignitas-articles/biotechnology-human-enhancement-and-the-end-of-medicine.

[3]Juengst, Eric, and Daniel Moseley. "Human Enhancement (Stanford Encyclopedia of Philosophy)." Stanford.Edu, 2015, https://plato.stanford.edu/entries/enhancement/. Accessed 25 Aug. 2026.

[4]Ostberg, René. "Transhumanism | Definition, Origins, Characteristics, & Facts | Britannica." Encyclopædia Britannica, 2019, www.britannica.com/topic/transhumanism.

[5]Dvorsky, George. "What Ever Happened to the Transhumanists?" Tech News. Gizmodo, 2022, https://gizmodo.com/what-happened-to-transhumanism-in-2022-life-extension-1849199492.

[6]Fukuyama, Francis. "Transhumanism." Foreign Policy, no. 144, Washingtonpost.Newsweek Interactive, LLC, Sept.-Oct. 2004, pp. 42-43. JSTOR, http://www.jstor.org/stable/4152980. Accessed 2 May. 2026.

[7]"Fast Facts: 1 In 4 People Globally Still Lack Access to Safe Drinking Water – WHO, UNICEF." Unicef.Org, 2025, https://www.unicef.org/press-releases/fast-facts-1-4-people-globally-still-lack-access-safe-drinking-water-who-unicef. Accessed 25 Aug. 2026.

[8]"Our Mission." Mission, Humanity+, 02 May 2026, www.humanityplus.org/about.
[9]Kass, Leon R. "Beyond Therapy: Biotechnology and the Pursuit of Human Improvement." President's Council on Bioethics, Jan. 2003. https://bioethicsarchive.georgetown.edu/pcbe/background/kasspaper.html. Accessed 7 Mar. 2025.
[10]Langdon Winner. The Whale and the Reactor : A Search For Limits In An Age Of High Technology. University Of Chicago Press, 1986, pp. 19–39.
[11]Jia Tolentino, "The Age of Instagram Face," Trick Mirror: Reflections on Self-Delusion (Random House, 2019).
[12]Donna Jeanne Haraway. A Cyborg Manifesto : Science, Technology, and Socialist-Feminism in the Late Twentieth Century. 2009.
[13]Sloan, MIT. "'AI Gravity' Is Pulling You toward Dependency. Here's How to Push Back | MIT Sloan." MIT Sloan, 2026, https://mitsloan.mit.edu/ideas-made-to-matter/ai-gravity-pulling-you-toward-dependency-heres-how-to-push-back. Accessed 25 Aug. 2026.
[14]Manuel Gil Martins, Henrique. "The Upper Limit: An Essay on Mental Integrity and Mental Enhancement." Medicine, Law & Society, vol. 15, no. 2, University of Maribor, 2022, https://doi.org/10.18690/mls.15.2.217-240.2022. Accessed 24 Sept. 2026.
[15]Gilbert, Frederic, et al. "I Miss Being Me: Phenomenological Effects of Deep Brain Stimulation." AJOB Neuroscience, vol. 8, no. 2, Informa UK Limited, Apr. 2017, pp. 96–109, https://doi.org/10.1080/21507740.2017.1320319.
[16]Ienca, Marcello, and Roberto Andorno. "Towards New Human Rights in the Age of Neuroscience and Neurotechnology." Life Sciences, Society and Policy, vol. 13, no. 1, Apr. 2017, https://doi.org/10.1186/s40504-017-0050-1. Accessed 24 Sept. 2026.
[17]Kraut, Richard. "Aristotle's Ethics." Stanford Encyclopedia of Philosophy, 2 July 2022, https://plato.stanford.edu/entries/aristotle-ethics/. Accessed 9 Sept. 2026.
[18]Sartre, Jean-Paul. Existentialism Is a Humanism. 1946. Translated by Carol Macomber, Yale University Press, 2007.
[19]Aronson, Ronald. "Albert Camus." Stanford Encyclopedia of Philosophy, Stanford University, 27 Oct. 2011, https://plato.stanford.edu/entries/camus/. Accessed 9 Sept. 2026.
[20]Lilley, Stephen. "Introduction to the Transhumanity Debate." Sociology Faculty Publications, Sacred Heart University, 2012. https://digitalcommons.sacredheart.edu/sociol_fac.
[21]World Health Organization. "Tobacco." World Health Organization, World Health Organization, 25 June 2025, https://www.who.int/news-room/fact-sheets/detail/tobacco. Accessed 25 Aug. 2026.
[22]The Editors of Encyclopaedia Britannica. "Russian Revolution." Encyclopædia Britannica, 5 Mar. 2024, https://www.britannica.com/event/Russian-Revolution. Accessed 9 Sept. 2026.
[23]Hartgens, F., Kuipers, H. Effects of Androgenic-Anabolic Steroids in Athletes. Sports Med 34, 513–554 (2004). https://doi.org/10.2165/00007256-200434080-00003.
[24]Anna Wexler, A pragmatic analysis of the regulation of consumer transcranial direct current stimulation (TDCS) devices in the United States, Journal of Law and the Biosciences, Volume 2, Issue 3, November 2015, Pages 669–696, https://doi.org/10.1093/jlb/lsv039.

The Access Paradox: Who Gets to Benefit When Medicine Becomes a Market

Solenyi Onwume[1]
[1]*Atholton High School, Maryland, Howard County*

Abstract

While medical advancements have provided humanity with the ability to treat and cure diseases once marked fatal, continuing mortality rates prove that a treatment's existence is not synonymous with its accessibility. Regulatory bodies like the FDA compensate pharmaceutical companies for the extraordinary costs associated with pharmaceutical development, both pre- and post- drug approval. However, post-approval compensation often grants these companies immense power over pricing and production, leading to a profit-based monopoly system that leaves consumers unable to afford existing treatments. Medicine is both a product of significant financial and scientific investment and, as recognized by international frameworks such as the WHO constitution, something people have a fundamental right to access.[1] This dual identity introduces an ethical dilemma: how can society balance the incentivization of pharmaceutical innovation with the moral obligation of making lifesaving medicine accessible? This paper investigates the context used to justify drug industry incentivization and its implications for patient care. It proposes three principles which regulatory bodies must consider when approaching the issue of pharmaceutical incentivization. It further suggests policy changes which should be imposed on the drug industry to make progress towards ameliorating what this paper terms the innovation-access paradox.

1. The Paradox of Modern Medicine

Over the past century, global medical advancements have demonstrated humanity's capacity to overcome seemingly intractable diseases. In the modern era, diseases like tuberculosis, HIV, and malaria are now treatable.[2] Yet, these diseases continue to kill millions every year. These deaths are the consequence of a tenuous paradox: a treatment's existence is not synonymous with its accessibility. Nearly 2 billion people worldwide lack consistent access to necessary, existing medications.[3] Medicine is both a product of significant financial and scientific investment, and something international organizations assert people have a fundamental right to access.[1] This dual identity produces a core ethical conflict: how can society balance the incentivization of pharmaceutical innovation with the moral obligation of making lifesaving medicine accessible?

2. The Cost of Getting It Right

Pharmaceutical innovation requires extraordinary cost and risk. Producers face lengthy development processes, high failure rates, clinical trials, regulatory requirements, and manufacturing approvals. On average, it takes 10-15 years of research and 2.6 billion dollars to bring a new medication to market.[4] The process typically involves three steps:

1. Discovery and Preclinical Testing, during which potential compounds are identified and tested, typically takes 1-6 years.
2. Clinical trials, divided into three phases, take an average of 6-7 years, and

are crucial for evaluating efficacy and safety.
3. FDA applications, submitted post-trials for regulatory approval, can take anywhere from 6 months to two years depending on the submission type.[5]

Despite its tediousness, the long development process exists to prevent patients from facing the consequences of unsafe experimentation. Comprehensive regulations were introduced by the federal government as a protective measure following several pharmaceutical crises in the 20th century. Among the most notable of these crises was the thalidomide scandal of the late 1950s. Thalidomide was originally branded as a sedative for treatment of insomnia and anxiety. It was later marketed to pregnant women as an antiemetic, capable of quelling morning sickness. However, thalidomide soon revealed itself a powerful teratogen. The mass prescription of thalidomide led to more than ten thousand children experiencing lethal birth defects, including amelia, phomelia, and absence of bones due to exposure.[6] The tragedy led to FDA crackdowns. New regulations forced companies to disclose the side effects of new products.[7] The FDA also introduced the Drug Efficacy Study Implementation (DESI), to evaluate the safety of previously approved prescription drugs.[8]

The increased regulation and the inability to recover development costs could have led to fewer companies pursuing novel treatments. As a preventative measure, the government has introduced pre-approval economic assistance to lower daunting fiscal and procedural expenses. Research grants from government agencies like the National Institutes of Health (NIH), tax credits from specialized FDA programs, and public-private partnerships reduce risk.[9] While these protections benefit production, they also drive up costs for consumers. The existing drug development process demonstrates the indispensable trade-off that occurs where the pharmaceutical industry and its regulatory bodies interact. The same system that protects patient safety can lead to medication that is more costly and takes longer to produce.

Pharmaceutical Benefit Managers (PBMs) act as intermediaries in the pipeline from producer to consumer. PBMs negotiate with manufacturers to lower drug prices in exchange for favorable formulary placement, receiving rebates that make up the difference between the initial and lowered cost. These in turn impact out-of-pocket costs and the likelihood of prescription.[10]

3. When Incentives Encourage Inaccessibility

The government continues to reward drug production with post-approval economic incentives. Companies are granted exclusivity periods on drugs, typically 5 years long. These protections run concurrently with long-term patents, which last around 20 years. Together, they prevent cheaper generic competitors from entering the market, essentially manufacturing short-term monopolies. Direct-to-consumer (DTC) advertising also significantly drives up costs. Advertising pushes pricier, name-brand products onto patients while bypassing clinical skepticism. Expenses are passed onto patients, insurance providers, and physicians.[11] DTC advertising and exclusivity periods stunt the emergence of the healthy competition that balances efficacy and pricing in other industries.

These justifications are difficult to independently verify, as companies are not required to disclose actual development costs. A recent study found no correlation between estimated research and development costs and treatment costs.[12] Meanwhile, the U.S. government does not impose pricing controls on

pharmaceutical manufacturers, leaving them with immense pricing power. This leads to a profit-driven system, a structure which should draw particular ethical scrutiny in the medical industry. Company market power introduces skepticism as to whether drug pricing reflects fair compensation for innovation and resources, or is a guise to exploit patient dependence. Pharmaceutical companies frame high drug prices as compensation for research and development expenses, the clinical value of their treatment, and as funding for future innovation.[13]

Drug developers, patients, governmental bodies, and insurance companies must come to a consensus on the moral distinction between the commercialization of lifesaving drugs, and the commercialization of ordinary consumer products, and apply this distinction fairly to drug pricing and policy.

3a. Orphan Drugs

One clear paradigm of the innovation-access dilemma is orphan drug development. Orphan drugs are developed specifically for rare diseases and conditions, considered "orphans" because their patient population is too small for commercial viability.[14] Diseases are categorized as "orphans" if they affect less than 200,000 patients. However, over 7,000 of these rare diseases exist, with an estimated 1 in 10 Americans living with one of them.[15]

The small consumer base of orphan diseases leaves companies with little financial reason to develop treatments. As a result, governments provide extensive compensation through specialized grants, tax benefits, and regulatory pathways. The FDA's Orphan Drug Designation Program enables specialized benefits for drugs that treat rare diseases, including extended exclusivity periods.[16] Despite these benefits, orphan drugs are still typically 4.5 to 25 times more expensive than standard generically branded drugs, and individualized treatments can be up to 1000x more expensive.[17] One study published in 2023 found that amongst the drugs approved by the FDA from 2017 to 2021, the median drug cost for non-orphan drugs was 12,798 USD, while the median treatment cost for orphan drugs was 218,872 USD.[15] The case of orphan drugs not only demonstrates the role of regulatory bodies in encouraging the development of rare drug treatment, but also calls into question the obligation of these regulatory bodies to ensure patient affordability.

4. What the System Should Consider

Unlike most other industries, medicine occupies a morally atypical category in which company financial decisions directly impact whether someone lives or dies. The existing drug innovation-access paradigm presents three competing principles which must be considered in the context of pharmaceutical policy reform:

1. Innovation - companies deserve compensation for creating something valuable
2. Autonomy - as the creators, they deserve to have control over their intellectual property
3. Human Welfare - people have the right to receive medical care

5. Return to the Paradox

Modern medicine has reached a point where the question is no longer simply whether we *can* develop treatments, but *who* gets to benefit from them. The ethical dilemma of pharmaceutical innovation relies on how our perception of medicine, as a commodity or as a human right, places moral limits on the market.

To approach solutions to this dilemma, we suggest that regulatory bodies prioritize welfare over the autonomy of drug developers, a philosophy which already shapes policy in many countries around the world. A number of nations

like Mexico, France, and Turkey have successfully achieved significantly lower prescription drug prices than the United States.[20] Nations including Australia, Canada, and the United Kingdom, directly regulate affordability with value-based pricing and ban direct to consumer advertisement.[18,19]

Regulatory bodies in the United States should adopt similar policies by reducing post-approval benefits for drugs. Changes could include banning direct-to-consumer advertising, introducing value based pricing and lifting exclusivity periods to inhibit monopolization and put checks on pricing power. For value-based pricing systems, independent agencies evaluate clinical data to set medication prices based on the health benefits they offer in comparison to existing treatments. Value-based pricing could enforce fairness in a field otherwise motivated by production costs. This proposal bundles multiple mechanisms to facilitate incremental change that targets different aspects of the paradox - consumer engagement, exploitative pricing, and monopolization. To encourage drug development, especially for rare diseases with smaller patient bases, we believe that regulatory bodies should maintain or increase pre-approval benefits like tax breaks, research grants, and public-private partnerships. These benefits would motivate innovation without enabling exploitation by reforming how companies are incentivized.

6. Conclusion

This paper set out to explore ways in which society can balance the incentivization of pharmaceutical innovation with the moral obligation of making lifesaving medicine accessible. We conclude that collective welfare should be weighed above corporate autonomy, in a system where innovation is preserved through a different mechanism rather than abandoned. We propose a cluster of incremental changes to the current American drug production pipeline which would increase pre-approval benefits while cutting down on post-approval ones. At its core, the access paradox asks how much of a commodity medicine can remain before its status as a human necessity demands limits on the market. As the capabilities of healthcare advance, pharmaceutical companies and their regulators must work together to ensure ethics are not left behind. By introducing policies that inhibit monopolization and exploitation, we can make steady progress towards sealing the innovation-access gap.

7. References

[1]World Health Organization. "Constitution of the World Health Organization." *World Health Organization*, 1948, https://www.who.int/about/governance/constitution.

[2]"These 5 Diseases Are Treatable. They Kill Millions. | Partners in Health." *Www.Pih.Org*, 4 Aug. 2023, https://www.pih.org/article/these-5-diseases-are-treatable-they-kill-millions.

[3]Ozawa, Sachiko, et al. "Access to Medicines Through Health Systems in Low- and Middle-Income Countries." *Health Policy and Planning*, vol. 34, no. Supplement_3, Dec. 2019, pp. iii1–3, https://doi.org/10.1093/heapol/czz119.

[4]Ulrich, Jocelyn. "Research and Development Continues Long After a Medicine Is Initially Approved | PhRMA." *Phrma.Org*, 22 Feb. 2025, https://phrma.org/blog/research-and-development-continues-long-after-a-medicine-is-initially-approved.

[5]Roden, Brian. "The Staggering Cost of Drug Development: A Look at the Numbers –

GreenField Chemical Inc..” *Greenfield*, 10 Aug. 2023, https://greenfieldchemical.com/2023/08/10/the-staggering-cost-of-drug-development-a-look-at-the-numbers/.

[6]“Thalidomide Monograph for Professionals.” *Drugs.Com*, https://www.drugs.com/monograph/thalidomide.html. Accessed 25 Aug. 2026.

[7]Bren, L. “Frances Oldham Kelsey: FDA Medical Reviewer Leaves Her Mark on History.” *Permanent.Access.Gpo.Gov*, Mar. 2001, https://permanent.access.gpo.gov/lps1609/www.fda.gov/fdac/features/2001/201_kelsey.html.

[8]“Drug Efficacy Study Implementation (DESI).” *FDA*, Aug. 2020, https://www.fda.gov/drugs/enforcement-activities-fda/drug-efficacy-study-implementation-desi.

[9]“Economic Assistance and Incentives for Drug Development.” *FDA*, May 2026, https://www.fda.gov/drugs/cder-small-business-industry-assistance-sbia/economic-assistance-and-incentives-drug-development.

[10]Bollmeier, Suzanne G, and Scott Griggs. “The Role of Pharmacy Benefit Managers and Skyrocketing Cost of Medications.” *Missouri Medicine*, vol. 121, no. 5, Sept. 2024, p. 403, https://pmc.ncbi.nlm.nih.gov/articles/PMC11482839/.

[11]Parekh, Natasha, and William H. Shrank. “Dangers and Opportunities of Direct-to-Consumer Advertising.” *Journal of General Internal Medicine*, vol. 33, no. 5, 2018, pp. 586–87, https://doi.org/10.1007/s11606-018-4342-9.

[12]Wouters, Olivier J., et al. “Association of Research and Development Investments With Treatment Costs for New Drugs Approved From 2009 To 2018.” *JAMA Netw Open*, vol. 5, no. 9, Sept. 2022, p. e2218623, https://doi.org/10.1001/jamanetworkopen.2022.18623.

[13]Torreele, Els. “Why Are Our Medicines so Expensive? Spoiler: Not for the Reasons You Are Being Told….” *European Journal of General Practice*, vol. 30, no. 1, Feb. 2024, https://doi.org/10.1080/13814788.2024.2308006.

[14]“Orphan Drug - an Overview | ScienceDirect Topics.” *Www.Sciencedirect.Com*, https://www.sciencedirect.com/topics/medicine-and-dentistry/orphan-drug. Accessed 25 Aug. 2026.

[15]Althobaiti, Hana, et al. “Disentangling the Cost of Orphan Drugs Marketed in the United States.” *Healthcare*, vol. 11, no. 4, Feb. 2023, p. 558, https://doi.org/10.3390/healthcare11040558.

[16]U.S. Food and Drug Administration. “Designating an Orphan Product: Drugs and Biological Products.” *FDA*, 8 July 2022, https://www.fda.gov/industry/medical-products-rare-diseases-and-conditions/designating-orphan-product-drugs-and-biological-products.

[17]“New Study: Big Pharma Price-Gouging Medications for Rare Diseases at Staggering Rate.” *The Campaign for Sustainable RX Pricing*, 11 Sept. 2019, https://www.csrxp.org/new-study-big-pharma-price-gouging-medications-for-rare-diseases-at-staggering-rate/.

[18]Synott, Patricia. “A Value-Based Approach to America’s Costly Prescription Drug Problem.” *Www.Commonwealthfund.Org*, 6 May 2022, https://www.commonwealthfund.org/blog/2022/value-based-approach-americas-costly-prescription-drug-problem.

[19]Levitt, Aimee. “What Other Countries Could Teach the U.S. About Bringing Down Drug Prices.” *Stanford Graduate School of Business*, Stanford University, Nov. 2022, https://www.gsb.stanford.edu/insights/what-other-countries-could-teach-us-about-bringing-down-drug-prices.

[20]Buchholz, Katharina. “Infographic: U.S. Drug Prices Sky-High in International Comparison.”

Statista Infographics, 9 Aug. 2022,
https://www.statista.com/chart/27932/us-prescri
ption-drug-prices-in-international-comparison/.

Journal of Ethics in Scientific and Technological Innovations

The Ethics of Preserving Identity During Awake Craniotomy

Angel Zhang[1]
[1]*Carlucci American International School of Lisbon, Sintra, Portugal*

Abstract

Awake brain mapping during craniotomy is used to identify and determine which areas need to be preserved during tumor resection. During an awake craniotomy, the patient performs tasks while electrodes placed on the exposed cortex record neural activity through electrocorticography (ECoG). Neurosurgeons may also use direct electrical stimulation to briefly disrupt small cortical areas and, as resection progresses, underlying white matter pathways. If simulation causes the patient to falter, the stimulated tissue is identified as essential and preserved. This paper will first describe the scientific and technical basis of awake brain mapping. It then examines ethical implications. Particular attention is given to the onco-functional balance between maximizing tumor resection and preserving meaningful functioning, alongside ethical questions of autonomy, including the individualized preservation of pathways essential to each patient, informed consent under uncertainty, beneficence, and the risk of neurological harm. This discussion is especially relevant given the recent progress in connectome-guided surgery that emphasizes preserving critical cognitive function, as well as longitudinal outcome assessments that can track the results of such procedures. Overall, there has been a shift from preserving isolated neurological abilities toward preserving broader elements comprising personal identity, which brings the aforementioned ethical concerns. To maintain ethics in awake craniotomy, future research must improve task validity, patient-specific assessment, professional standards, and accessibility while avoiding the reduction of personal identity to measurable cognitive performance.

1. Introduction

Awake craniotomies are performed to treat brain tumors, epilepsy, and other neurological conditions.[1] These conditions are delicate to handle due to being located in the organ that is responsible for cognition, behavior, memory, and many of the capacities through which individuals understand and express themselves. During one, patients are awake for the functional-mapping stage and perform tasks. Electrodes placed on the exposed cortex record neural activity through electrocorticography (ECoG).[2] This allows neurological function to be assessed in real time while surgery is taking place. The procedure is commonly used when areas located near functionally important cortical or subcortical networks are being operated on. Surgeons may apply direct electrical stimulation (DES) to small areas of the exposed cortex and, as resection progresses, to underlying white-matter pathways. This is when brief electrical current transiently interferes with neuron signaling in the stimulated region.[3,4] If the patient struggles with the task, that tissue is considered essential and preserved.

Effects of such surgery can not only be measured by survival. There are differences between preserving basic function and preserving the broader capacities through which a person lives. Patients may remain physically independent while experiencing changes in identity or cognition that alter how they participate in relationships and previously meaningful activities. Research has shown that brain neoplasms and their treatment can affect

neurocognitive function as well as a patient's perception and expression of self.[5]

As neuroscience has shifted from a localizationist view of the brain toward hodotopic and connectomic models, emphasis is increasingly being placed on interconnected and interacting neural networks instead of isolated cortical regions.[6] This creates a less distinct boundary between preserving "basic" functions such as movement or speech and preserving higher-order capacities such as memory, emotional processing, social cognition, and decision-making. Because functions relevant to identity depend on multiple interconnected networks rather than a single identifiable region, it introduces complications into surgery to identify and preserve the pathways and regions needed for a specific patient.[7]

The ethical problems of preserving identity during awake craniotomy would be the onco-functional balance between maximizing tumor resection and preserving meaningful functioning, as well as questions concerning autonomy, informed consent under uncertainty, and beneficence.

2. Awake Craniotomy and Brain Mapping

2.1. Clinical Relevance

Diffuse low-grade gliomas. Diffuse lower-grade gliomas usually infiltrate surrounding tissue, which may lead to changes in structure and functions of each area.[8] As a result, even with imaging and based on anatomical structures, it is unclear which regions to preserve, leading to the need for awake mapping. In a longitudinal study of 157 patients with diffuse low-grade glioma undergoing awake mapping-based surgery, Lemaitre and colleagues found that 55.4% already had at least one measurable cognitive impairment before surgery. At three months, 86% of patients who underwent surgery showed no postoperative cognitive decline, including 10% who had improved.[9]

Selected high-grade gliomas. For high-grade gliomas, awake craniotomy is clinically valuable in selected patients with tumors located near important functional networks as real-time testing may allow the surgeon to distinguish tissue that can be removed from tissue whose disruption produces impairment.[10] This is supported by evidence from the international multicenter GLIOMAP cohort, which examined 1,047 patients undergoing resection of primary glioblastoma in eloquent locations. 134 awake cases were compared with 402 asleep resections and it was found that neurological deficits were less frequent after awake surgery at three months—22% versus 33% (p=0.019)—and at six months—26% versus 41% (p=0.0048). Median overall survival was 17.0 versus 14.0 months, and median progression-free survival was 9.0 versus 7.3 months.[11]

Epilepsy surgery. Awake mapping also has a long history in epilepsy surgery, particularly where an epileptogenic lesion or proposed resection lies close to the functionally important cortex. A 2024 matched study compared 15 patients undergoing awake epilepsy surgery with 30 patients treated with conventional asleep surgery. Although the difference in outcome was not statistically significant, four of the 15 awake patients had previously been judged inoperable because the epileptogenic region was too close to eloquent structures.[12] The study therefore suggests that awake mapping may sometimes expand operability.

Tumors involving language or motor networks. Language organization varies substantially between individuals and may be altered by tumor-related plasticity, as seen in Figure 1, involving interacting pathways

including the arcuate/superior longitudinal fascicular system (AF/SLF), inferior fronto-occipital fasciculus (IFOF), and inferior longitudinal systems (ILF).[13,14]

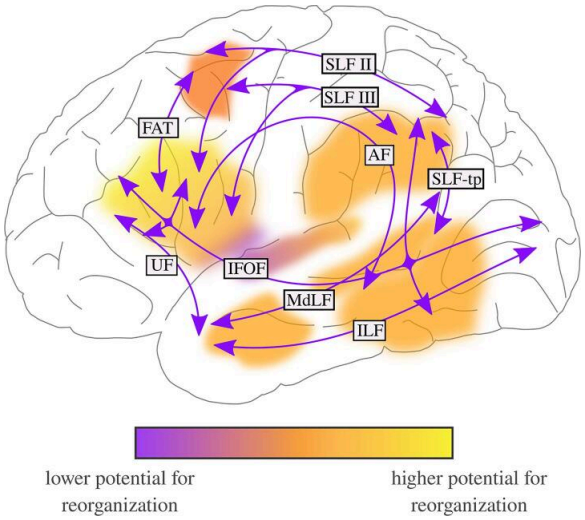


Figure 1. Schematic representation of cortical and subcortical components of the language network and their differing potential for functional reorganization. From Nieberlein et al. (2023).[13]

Motor mapping is also fundamental, particularly for lesions affecting the precentral region or corticospinal system. Sometimes, motor cortex and corticospinal pathways responsible for movement can be mapped and monitored electrophysiologically under general anesthesia.[15] The advantage of awake surgery is greater when the surgeons need to map functions, such as spatial attention or executive control, that cannot be accurately inferred from measuring motor evoked potentials, the electrical signals measured in muscles or the spinal cord after stimulating the motor cortex in the brain.[16]

Lesions near cortical or subcortical pathways. Awake brain mapping proves to be valuable when resection happens near subcortical pathways. In Duffau's study of 30 patients with left-hemisphere cortico-subcortical low-grade gliomas, stimulation of white matter pathways during resection produced disturbances in speech, dysarthria, and anomia. Fourteen resections were complete, whereas in 16 the tumor was not completely removed because functional pathways had been reached.[17]

2.2. Awake Brain Mapping

An asleep-awake-asleep technique is used frequently during awake craniotomy, where the patient is placed under general anesthesia during invasive parts like the scalp incision and skull opening, awakened during brain mapping and tumor removal, and put back to sleep for the final parts of the surgery.[18]

Preoperative assessment and task selection. The first stage is assessing the patient's functions prior to surgery to establish a baseline. This can identify deficits that already exist because of the tumor or epilepsy.[19] It also helps to determine if the patient would be able to perform a selected task accurately and consistently. In addition, De Witte and colleagues explicitly recommended tailoring intraoperative testing to the affected cortico-subcortical region as well as to the patient's occupation, hobbies, and everyday activities.[20]

Exposure and cortical mapping. After craniotomy and dural opening, the patient is woken for behavioral testing. DES is applied to exposed cortical sites while the patient performs tasks. Stimulation parameters are not completely standardized. In the 2024 review of executive-function mapping by Tariq et al., most included studies used bipolar stimulation at 60 Hz, with pulse durations of 1–2 ms and intensities of 2–6 mA.[21] Upon stimulation, if it causes an observable reaction, it is known as positive mapping. If there is no observable response, it suggests the area may be safer to

resect (remove) and is known as negative mapping.[22]

Postoperative assessment. Postoperative examination identifies immediate deficits in cognitive function while longitudinal studies use repeated assessment over weeks or months to assess impacts in the long term. Functional outcomes should also be interpreted at the level of real life, such as the patient's ability to return to work, their interpersonal functioning, independence, fatigue, and patient-reported cognitive change.[23]

2.3. Mapping Domains

Historically, awake mapping has focused on language and sensorimotor function as they have direct impacts on everyday life. In a recent review paper covering 272 publications from February 2017 through November 2023, de Zwart and Ruis found that 90.4% of studies reported language testing.[22] Notably, visuospatial functions, social cognition, and executive functions are receiving increasing attention.[16]

Executive functions. The 2024 systematic review by Tariq and colleagues identified 13 studies involving 351 patients in whom executive functions were assessed intraoperatively. Reported paradigms included the Stroop task, spatial 2-back tasks, line bisection, trail-making tasks, and digit span.[21] The review concluded that executive-function mapping is feasible and reported favorable outcomes.

Social cognition. Social cognition includes the processes used to infer and respond to other people's emotions, beliefs, intentions, and perspectives. One component frequently studied in neurosurgical research is mentalizing or theory of mind—the capacity to infer another person's mental state.[24] Awake stimulation studies in patients with right-sided diffuse low-grade gliomas have provided evidence that these processes depend on distributed cortical-subcortical networks, emphasizing its importance as a subject of brain mapping.

Memory. De Zwart and Ruis identified memory, together with executive functions, as a domain in which further development of intraoperative tests remains necessary since standard DES perturbs processing over a very short period and can not test for memory processes like encoding, consolidation, retention, and retrieval, which happens over longer time intervals.[22]

Skilled and occupation-specific functions. Increasing interest in personalized mapping has led some teams to adapt intraoperative testing to professional or personal skills. One well-described example involved a professional violinist for whom motor and musical performance were incorporated into the awake procedure.[25]

3. Ethical Implications

3.1. The Onco-Functional Balance

For diffuse gliomas, particularly slow-growing lower-grade tumors, it is hard to fully pinpoint the boundary between tumor and eloquent brain. As gliomas infiltrate surrounding tissue over time, functional networks may reorganize through neuroplasticity.[26, 27, 28] This reorganization can occur during tumor growth and continue after surgery, meaning that functional organization may differ between patients and may also change within the same patient over time. This creates an ethical problem where even with brain mapping, it is hard to determine how far resection should proceed.

The main ethical questions arise from the issue of the onco-functional balance.[26]

Greater extent of resection in diffuse lower-grade glioma has repeatedly been associated with improved tumor control and prognosis. Leaving safely removable tumor mass cannot be regarded as a neutral choice. Clinical practice therefore aims for maximal safe resection. This comes with functional risks. Permanent surgically acquired deficits can reduce independence and quality of life and may interfere with subsequent oncological treatment. Balancing these factors means that surgery should seek the greatest reasonable oncological benefit while remaining within a level of functional risk that the informed patient considers acceptable, making the decision more individualized.[29] Intraoperative mapping provides a means of doing so by identifying patient-specific functional boundaries while resection is occurring.

3.2. Autonomy & Informed Consent Under Uncertainty

Predicting possible postoperative identity changes is difficult, which is why patients need to make an informed decision and consent to the uncertainty of awake craniotomy. Patients must understand the trade-off at the point at which further tumor removal may produce additional oncological benefit while increasing the probability of functional damage. Ethical analyses of awake craniotomy have identified informed consent as a central concern, underscoring that patients need clear information about both tumor behavior and the expected medical and social consequences of surgery.[30]

This information comes with uncertainty due to neuroplasticity and individual functional variability as mentioned in Section 3.1. Neither imaging nor stimulation provides comprehensive information about all human capacities. Comparisons of fMRI and transcranial magnetic stimulation with cortical stimulation have also demonstrated variable accuracy.[31] Current informed consent already addresses the risks associated with tumor resection. However, ethically meaningful consent should also make explicit the limitations of functional mapping itself. Patients should understand that a function may be affected despite negative mapping and preserving performance on one task does not guarantee preservation of the broader ability. Consent should communicate that stopping resection may reduce neurological risk but leave more tumor whilst continuing resection may improve oncological outcomes while increasing functional risk.[26]

Informed consent should therefore communicate uncertainty itself. This allows for autonomy, which has a substantive, values-based component.[32] Patients may rationally prioritize outcomes differently. For example, one person may accept a higher risk of subtle cognitive impairment to maximize tumor removal; another may regard preservation of a particular professional skill or ability as more important.

3.3. Beneficence and Quality of Life

The principle of beneficence requires clinicians to promote the patient's welfare, both in terms of controlling disease and ensuring quality of life.[33] Preventable cognitive or physical deficits may impact identity, making their avoidance ethically important. However, disability should not automatically be equated with poor quality of life. The so-called disability paradox shows that people frequently underestimate the quality of life reported by people living with disabilities. Albrecht and Devlieger's classic study found that many participants with serious and persistent disability nevertheless described their quality of life positively.[34]

Ethical decision-making must therefore avoid two extremes, one of underestimating the consequences of preventable neurological injury

and the other that assuming that disability necessarily destroys well-being or identity, and prioritize minimizing avoidable changes that would substantially interfere with the life that the patient considers important. Patients should therefore also be educated on this paradox in order to make a fully informed decision about where they want their own onco-functional balance to be.

4. Current Progress

Current progress in awake glioma surgery is increasingly directed toward personalized preservation of cognition and identity-related functions.

4.1. Multiparadigm Mapping

Intraoperative mapping is gradually expanding functions that it is able to test for. The proportion of studies assessing only one cognitive domain decreased, suggesting a shift toward multiparadigm mapping. However, higher-order mapping remains less standardized than motor and language mapping, and some tasks lack sufficient validation for routine clinical use.[16]

4.2. Integration of Neuropsychology

The role of neuropsychological assessment before and after surgery is becoming more important.[35] Preoperative testing establishes an individual baseline by profiling cognitive and psychosocial functionings of a patient. It can also guide which functions should be tested intraoperatively as well as help with postoperative rehabilitation.[36]

4.3. Connectome-Guided Surgery

Connectome-guided surgery treats the brain as a set of interconnected functional networks.[37]

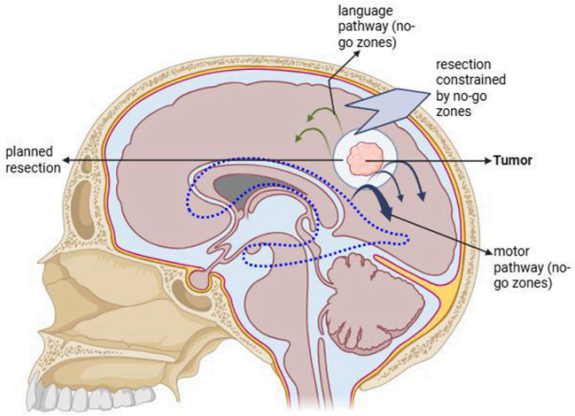


Figure 2. Patient example of tumor resection constrained by connectome-informed no-go zones.[37]

Surgical planning therefore aims to preserve both important cortical areas and the white-matter pathways that connect them.[38] Preoperative tools such as tractography and functional MRI can help identify networks that may be at risk and guide where intraoperative testing should focus. These techniques complement DES as they can not identify the function of each region. Tractography estimates the probable course of white-matter pathways without showing whether they play any key functional roles, while functional imaging can be affected by changes caused by a tumor and individual variability.[39]

4.4. Longitudinal Assessment

There is also increasing emphasis on assessing patients over months and years through longitudinal assessments. With neuroplasticity, functions and outcomes can recover or deteriorate over time. In addition it allows for the measurement of long term impacts on patients' lives. For example, a 2026 series of 515 employed patients with IDH-mutant grade 2 glioma undergoing awake mapping-guided surgery reported that 91.5% returned to work within one year. Such data illustrate the growing

effort to measure whether patients can resume meaningful lives.[40]

5. Conclusion

Awake craniotomy and brain mapping has evolved from solely focusing on motor and language function to higher order functions, with an emphasis on being patient-specific. The ethical balance between maximizing tumor removal and minimizing neurological changes that may affect quality of life, autonomy, or identity is also being increasingly recognized. Because neuroplasticity and individual variability mean that functional organization cannot be predicted reliably from anatomy or imaging alone, direct electrical stimulation remains important for identifying functional boundaries during surgery. Ethical practice demands individualized mapping, transparent consent about uncertainty, and attention to the functions each patient values most. Future work should focus on validating higher-order cognitive tasks, incorporating neuropsychological and connectome-based information, establishing clearer professional standards, and improving access to these techniques.

6. References

[1]KRAMBEK, MÁRCIO CARDOSO, et al. "Awake Craniotomy in Brain Tumors - Technique Systematization and the State of the Art." *Revista Do Colégio Brasileiro de Cirurgiões*, vol. 48, FapUNIFESP (SciELO), 2021, https://doi.org/10.1590/0100-6991e-20202722. Accessed 26 Aug. 2026.

[2]Safa, Adrian, et al. "Intraoperative Electrocorticography in Awake Brain Surgery: Comparing Three Electrode Configurations for Detecting Stimulus-Induced and Spontaneous Epileptiform Activity." *Journal of Neurosurgery*, vol. 144, no. 3, Journal of Neurosurgery Publishing Group (JNSPG), Mar. 2026, pp. 570–78, https://doi.org/10.3171/2025.7.jns251224. Accessed 26 Aug. 2026.

[3]Szelényi, Andrea, et al. "Intraoperative Electrical Stimulation in Awake Craniotomy: Methodological Aspects of Current Practice." *Neurosurgical Focus*, vol. 28, no. 2, Journal of Neurosurgery Publishing Group (JNSPG), Feb. 2010, p. E7, https://doi.org/10.3171/2009.12.focus09237. Accessed 26 Aug. 2026.

[4]Mahon, Bradford Z., et al. "Direct Electrical Stimulation Mapping of Cognitive Functions in the Human Brain." *Cognitive Neuropsychology*, vol. 36, no. 3–4, Informa UK Limited, May 2019, pp. 97–102, https://doi.org/10.1080/02643294.2019.1630375 . Accessed 26 Aug. 2026.

[5]Anderson-Shaw, Lisa, et al. "Brain Neoplasm and the Potential Impact on Self-Identity." *AJOB Neuroscience*, vol. 1, no. 3, Informa UK Limited, June 2010, pp. 3–7, https://doi.org/10.1080/21507740.2010.485122. Accessed 24 Aug. 2026.

[6]Roy, Arnab, et al. "From Localization to Connectomics: A Contemporary View of Human Brain Structure and Dynamic Network Function." *International Journal of Scientific Research and Technology*, Oct. 2025, pp. 483–93, https://doi.org/10.5281/zenodo.17455149. Accessed 26 Aug. 2026.

[7]De Roeck, Laurien, et al. "Increased Functional Network Segregation in Glioma Patients Posttherapy: A Neurological Compensatory Response or Catastrophe for Cognition?" *Network Neuroscience*, vol. 9, no. 2, June 2025, p. 743, https://pmc.ncbi.nlm.nih.gov/articles/PMC12226146/. Accessed 26 Aug. 2026.

[8]Duffau, Hugues. "Mapping the Connectome in Awake Surgery for Gliomas: An Update." *Journal of Neurosurgical Sciences*, vol. 61, no. 6, Edizioni Minerva Medica, Sept. 2017,

https://doi.org/10.23736/s0390-5616.17.04017-6. Accessed 26 Aug. 2026.
[9]Lemaitre, Anne-Laure et al. "Cognitive preservation following awake mapping-based neurosurgery for low-grade gliomas: A longitudinal, within-patient design study." *Neuro-oncology* vol. 24,5 (2022): 781-793. doi:10.1093/neuonc/noab275. Accessed 26 Aug. 2026.
[10]Pallud, J., et al. "Direct Electrical Bipolar Electrostimulation for Functional Cortical and Subcortical Cerebral Mapping in Awake Craniotomy. Practical Considerations." *Neurochirurgie*, vol. 63, no. 3, Elsevier BV, June 2017, pp. 164–74, https://doi.org/10.1016/j.neuchi.2016.08.009. Accessed 26 Aug. 2026.
[11]Gerritsen, Jasper K. W., et al. "Effect of Awake Craniotomy in Glioblastoma in Eloquent Areas (GLIOMAP): A Propensity Score-Matched Analysis of an International, Multicentre, Cohort Study." *The Lancet Oncology*, vol. 23, no. 6, Elsevier BV, June 2022, pp. 802–17, https://doi.org/10.1016/s1470-2045(22)00213-3. Accessed 26 Aug. 2026.
[12]Vigren, Patrick, et al. "Awake Craniotomy in Epilepsy Surgery Includes Previously Inoperable Patients with Preserved Efficiency and Safety." *International Journal of Neuroscience*, vol. 134, no. 12, Informa UK Limited, Nov. 2023, pp. 1624–29, https://doi.org/10.1080/00207454.2023.2279498. Accessed 26 Aug. 2026.
[13]Nieberlein, Laura, et al. "Reorganization and Plasticity of the Language Network in Patients with Cerebral Gliomas." *NeuroImage: Clinical*, vol. 37, Elsevier BV, 2023, p. 103326, https://doi.org/10.1016/j.nicl.2023.103326. Accessed 26 Aug. 2026.
[14]Conner, Andrew K, et al. "A Connectomic Atlas of the Human Cerebrum—Chapter 10: Tractographic Description of the Superior Longitudinal Fasciculus." *Operative Neurosurgery*, vol. 15, no. Suppl 1, Sept. 2018, p. S407, https://pmc.ncbi.nlm.nih.gov/articles/PMC6887962/. Accessed 26 Aug. 2026.
[15]Morshed, Ramin A., et al. "Functional Mapping for Glioma Surgery, Part 2." *Neurosurgery Clinics of North America*, vol. 32, no. 1, Elsevier BV, Jan. 2021, pp. 75–81, https://doi.org/10.1016/j.nec.2020.09.001. Accessed 26 Aug. 2026.
[16]de Zwart, Beleke, and Carla Ruis. "An Update on Tests Used for Intraoperative Monitoring of Cognition during Awake Craniotomy." *Acta Neurochirurgica*, vol. 166, no. 1, Springer Science and Business Media LLC, May 2024, https://doi.org/10.1007/s00701-024-06062-6. Accessed 26 Aug. 2026.
[17]Duffau, Hugues et al. "Intraoperative mapping of the subcortical language pathways using direct stimulations. An anatomo-functional study." *Brain : a journal of neurology* vol. 125,Pt 1 (2002): 199-214. doi:10.1093/brain/awf016
[18]Huncke, Kate, et al. "The Asleep-Awake-Asleep Anesthetic Technique for Intraoperative Language Mapping." *Neurosurgery*, vol. 42, no. 6, Ovid Technologies (Wolters Kluwer Health), June 1998, pp. 1312–16, https://doi.org/10.1097/00006123-199806000-00069. Accessed 26 Aug. 2026.
[19]Aabedi, Alexander A., et al. "Assessment of Wakefulness during Awake Craniotomy to Predict Intraoperative Language Performance." *Journal of Neurosurgery*, vol. 132, no. 6, Journal of Neurosurgery Publishing Group (JNSPG), June 2020, pp. 1930–37, https://doi.org/10.3171/2019.2.jns183486. Accessed 26 Aug. 2026.
[20]De Witte, E., et al. "Subcortical Language and Non-Language Mapping in Awake Brain Surgery: The Use of Multimodal Tests." *Acta Neurochirurgica*, vol. 157, no. 4, Springer Science and Business Media LLC, Jan. 2015, pp. 577–88,

https://doi.org/10.1007/s00701-014-2317-0. Accessed 26 Aug. 2026.
[21]Tariq, Rabeet, et al. "Intraoperative Mapping and Preservation of Executive Functions in Awake Craniotomy: A Systematic Review." *Neurological Sciences*, vol. 45, no. 8, Springer Science and Business Media LLC, Mar. 2024, pp. 3723–35, https://doi.org/10.1007/s10072-024-07475-y. Accessed 26 Aug. 2026.
[22]de Zwart, Beleke, and Carla Ruis. "An Update on Tests Used for Intraoperative Monitoring of Cognition during Awake Craniotomy." *Acta Neurochirurgica*, vol. 166, no. 1, Springer Science and Business Media LLC, May 2024, https://doi.org/10.1007/s00701-024-06062-6. Accessed 26 Aug. 2026.
[23]Lemaitre, Anne-Laure et al. "Cognitive preservation following awake mapping-based neurosurgery for low-grade gliomas: A longitudinal, within-patient design study." *Neuro-oncology* vol. 24,5 (2022): 781-793. doi:10.1093/neuonc/noab275
[24]Yordanova, Yordanka Nikolova et al. "Neural pathways subserving face-based mentalizing." *Brain structure & function* vol. 222,7 (2017): 3087-3105. doi:10.1007/s00429-017-1388-0
[25]Piai, Vitória et al. "Awake Surgery for a Violin Player: Monitoring Motor and Music Performance, A Case Report." *Archives of clinical neuropsychology : the official journal of the National Academy of Neuropsychologists* vol. 34,1 (2019): 132-137. doi:10.1093/arclin/acy009
[26]Duffau, Hugues, and Emmanuel Mandonnet. "The 'Onco-Functional Balance' in Surgery for Diffuse Low-Grade Glioma: Integrating the Extent of Resection with Quality of Life." *Acta Neurochirurgica*, vol. 155, no. 6, Springer Science and Business Media LLC, Feb. 2013, pp. 951–57, https://doi.org/10.1007/s00701-013-1653-9. Accessed 26 Aug. 2026.
[27]Ng, Sam, et al. *Distributed under a Creative Commons CC BY 4.0 -Attribution -International License Perspectives in Human Brain Plasticity Sparked by Glioma Invasion: From Intraoperative (Re)Mappings to Neural Reconfigurations*. https://doi.org/10.4103/1673-5374.382246⟩.
[28]Duffau, Hugues. "Lessons from brain mapping in surgery for low-grade glioma: insights into associations between tumour and brain plasticity." *The Lancet. Neurology* vol. 4,8 (2005): 476-86. doi:10.1016/S1474-4422(05)70140-X
[29]Aaronson, Neil K., et al. "Compromised Health-Related Quality of Life in Patients With Low-Grade Glioma." *Journal of Clinical Oncology*, vol. 29, no. 33, American Society of Clinical Oncology (ASCO), Nov. 2011, pp. 4430–35, https://doi.org/10.1200/jco.2011.35.5750. Accessed 26 Aug. 2026.
[30]Kirsch, Brandon, and Mark Bernstein. "Ethical Challenges with Awake Craniotomy for Tumor." *Canadian Journal of Neurological Sciences / Journal Canadien Des Sciences Neurologiques*, vol. 39, no. 1, Cambridge University Press (CUP), Jan. 2012, pp. 78–82, https://doi.org/10.1017/s0317167100012737. Accessed 26 Aug. 2026.
[31]Weng, Hsu-Huei et al. "Accuracy of Presurgical Functional MR Imaging for Language Mapping of Brain Tumors: A Systematic Review and Meta-Analysis." *Radiology* vol. 286,2 (2018): 512-523. doi:10.1148/radiol.2017162971
[32]Hewins, Will, et al. "The Effects of Brain Tumours upon Medical Decision-Making Capacity." *Current Oncology Reports*, vol. 21, no. 6, Springer Science and Business Media LLC, May 2019, https://doi.org/10.1007/s11912-019-0793-3. Accessed 26 Aug. 2026.
[33]Varkey, Basil. "Principles of Clinical Ethics and Their Application to Practice." *Medical*

Principles and Practice, vol. 30, no. 1, S. Karger AG, June 2020, pp. 17–28, https://doi.org/10.1159/000509119. Accessed 26 Aug. 2026.
[34]Albrecht, Gary L., and Patrick J. Devlieger. "The Disability Paradox: High Quality of Life against All Odds." *Social Science & Medicine*, vol. 48, no. 8, Elsevier BV, Apr. 1999, pp. 977–88, https://doi.org/10.1016/s0277-9536(98)00411-0. Accessed 26 Aug. 2026.
[35]Weller, Michael, et al. "EANO Guidelines on the Diagnosis and Treatment of Diffuse Gliomas of Adulthood." *Nature Reviews Clinical Oncology*, vol. 18, no. 3, Springer Science and Business Media LLC, Dec. 2020, pp. 170–86, https://doi.org/10.1038/s41571-020-00447-z. Accessed 26 Aug. 2026.
[36]Hande, Vasudha H, et al. "Role of Clinical Neuropsychologists in Awake-Craniotomy." *Neurology India*, vol. 69, no. 3, Ovid Technologies (Wolters Kluwer Health), May 2021, pp. 711–16, https://doi.org/10.4103/0028-3886.319237. Accessed 26 Aug. 2026.
[37]Patel, Tirath, et al. "Emerging Paradigms in Connectome-Guided Neurosurgery for Central Nervous System Tumors: A Narrative Review of Structural and Functional Connectomics." *Annals of Medicine & Surgery*, vol. 88, no. 6, Ovid Technologies (Wolters Kluwer Health), Apr. 2026, pp. 3276–83, https://doi.org/10.1097/ms9.0000000000004915. Accessed 26 Aug. 2026.
[38]Ghinda, Cristina Diana, and Hugues Duffau. "Network Plasticity and Intraoperative Mapping for Personalized Multimodal Management of Diffuse Low-Grade Gliomas." *Frontiers in Surgery*, vol. 4, Frontiers Media SA, Jan. 2017, https://doi.org/10.3389/fsurg.2017.00003. Accessed 26 Aug. 2026.
[39]Voets, Natalie L., et al. "Diffusion Tractography for Awake Craniotomy: Accuracy and Factors Affecting Specificity." *Journal of Neuro-Oncology*, vol. 153, no. 3, Springer Science and Business Media LLC, July 2021, pp. 547–57, https://doi.org/10.1007/s11060-021-03795-7. Accessed 26 Aug. 2026.
[40]Duffau, Hugues. "Correlation Between Resumption of Employment and Overall Survival Following Awake Mapping–Based Surgery for IDH-Mutant Grade 2 Glioma: A Consecutive Series With 515 Patients." *Journal of Neurosurgery*, vol. 2, 2026, pp. 1–10, https://doi.org/10.3171/2026.2.JNS252614.

Editing the Self: Personal Identity in Therapeutic Memory Modification

Eunsoo Lily Oh[1*]
[1]*St. Mark's School, Massachusetts, USA*

Abstract

Treatments designed to address traumatic memories while they are in the process of reconsolidation are currently under close examination. However, there is still only preliminary data to show that these treatments consistently provide a long-term reduction in symptoms. This raises a significant point: if changing a traumatic memory could change a person's sense of self, then how should we define what counts as such a change? To answer this, one must distinguish between a person's continuing existence and shifts in their personal commitments or how they understand their life. This paper compares John Locke's ideas on personal identity based on consciousness, Joseph Butler's challenge to grounding identity in memory, Derek Parfit's concepts of psychological connection and continuity, and Paul Ricoeur's views on narrative identity. It suggests that Parfit offers the most useful starting point: a person's link to their past includes intentions, attachments, and beliefs, not just what they remember happening. Narrative identity builds on this by explaining how these connections help people make sense of who they are. The paper highlights two potential issues: existing psychological links can be mistaken for an unchanging bond to the past, and a familiar way of describing oneself can be mistaken for proof that the connections it describes are still present. Using both perspectives leads to distinct questions about what a patient still does and about what those actions now mean to them, treating any difference between the answers as something that needs explanation rather than being ignored.

1. What Does It Mean to Change Someone's Identity?

Changing memories through therapy brings up a seemingly simple question: if our memories define us, then altering them could alter who we are. However, the significance of this question relies on how we define "who we are." Someone who quits a profession or shifts their religious views might state, "I'm not the same person I once was." Typically, this indicates a change in their values, not that one individual has vanished and a different one has emerged. This same difference is important when discussing changes to memory.

Philosophers refer to remaining the exact same individual over time as "numerical identity." This is distinct from keeping the same characteristics. A person can stay the same individual while becoming more confident, less afraid, or less committed to a previous aim. Therefore, worries about memory modification might refer to a significant change in someone's life without suggesting the person literally disappears. This paper focuses on that psychological worry: what connects a patient to their past experiences, and what would it mean for treatment to change those connections? Research offers a reason to ask this question, though its implications need careful explanation. Nader and colleagues demonstrated that reactivated fear memories in rats could become open to disruption during reconsolidation, the process where a retrieved memory becomes stable again.[1] Brunet and colleagues later studied propranolol after traumatic memory retrieval in patients with post-traumatic stress disorder and reported weaker physiological responses during subsequent traumatic imagery.[2] These findings

do not prove that complex life memories can be selectively erased. They do make it relevant, however, to distinguish remembering what happened from still reacting to it in the same way.

This paper argues that Parfit's explanation of psychological connection and continuity should be the starting point for understanding identity concerns about these treatments. It shows how a person stays connected to the past through more than direct recall. Ricoeur's narrative explanation should supplement it, because simply having a connection does not fully explain its importance to the person. Both are useful because these findings can separate: a commitment can continue even if its meaning changes, and a familiar explanation of that commitment can remain after the attachment itself has weakened. In practice, using both helps a clinician ask different questions of a patient, which Section 7 outlines. The argument addresses which explanation best clarifies the identity question, rather than whether a particular treatment should be allowed.

2. Why Modify Memory at all?

PTSD is common enough to be a genuine problem, not just an abstract concept. About 7.8% of American civilians will develop it at some point in their lives, with women affected almost twice as often as men.[3] This number increases considerably among veterans getting VA care; for individuals who served in conflicts after September 11, PTSD diagnoses are as high as 24%.[4] Sexual assault survivors also experience an even greater occurrence of the condition. Worldwide, data from the World Mental Health Surveys indicates a lifetime prevalence of around 3.9%.[5] For people with PTSD, past incidents are more than just stored memories; they provoke flashbacks, increased vigilance, and avoidance behaviors that are severe enough to disrupt daily life, sometimes lasting many years after the original traumatic event.

Exposure therapy, the standard treatment, works by weakening connections. Repeated, safe encounters with the memory build a new link that competes with the initial fear response, rather than removing it. Because the first connection stays, it can reappear under stress, a phenomenon called reinstatement.[6]

Reconsolidation-based methods aim to do something different from extinction therapy: they target the original memory trace itself during the short time it becomes unstable after being recalled, instead of simply layering a new memory over it.

Evidence for this method exists, but it is still developing. A meta-analysis looking at propranolol's effects on emotional memory showed a moderate decrease in recalling negative information later on when the drug was given during reconsolidation. However, this estimate comes from only eight studies involving healthy adults, not patients with clinical PTSD, and the authors asked for more research to see if the effect would last in traumatized populations.[7] One widely cited behavioral technique, which involves updating a fear memory with new information during the reconsolidation window, was first reported to prevent the return of fear across several standard tests.[8] Yet a later verification report, which reanalyzed the original data using more careful criteria for excluding participants, could not confirm the reported effect.[9] The clinical finding by Brunet and colleagues discussed in Section 1 also came from a small sample.[2]

Taken together, this research does not confirm that reconsolidation-based treatment reliably and lastingly reduces PTSD symptoms. What it does show is that such a treatment represents an active and studied possibility, rather than a purely theoretical one. At least some evidence, though provisional and disputed, points to a decrease in a particular kind of suffering: the

physical and behavioral reaction to a memory, rather than the memory's content itself.
This is why the question of identity, which this paper addresses, is not just an academic point. If such treatments prove effective, patients and clinicians will need to balance possible relief from documented suffering against concerns that treatment could change something the patient values.[10] This paper does not decide how that balance should be made, and the uncertain evidence is one reason not to. It addresses an earlier question: what an identity-based concern would need to specify, what would be lost, and under what circumstances, when a memory's emotional hold changes rather than its content. This is precisely the distinction the remainder of this paper explores.

3. Locke: Recognizing a Past Experience as My Own

Locke believed that personal identity comes from a person's consciousness reaching back to include previous thoughts and actions.[11] To grasp this concept, consider the difference between recalling your own first day at school and simply knowing that someone else went to that school. Both situations involve information from the past. Yet, only the first presents an earlier event as something you personally experienced. Locke placed this first-person connection at the core of his explanation for personal identity: our current awareness recognizes an earlier experience as its own.

This goes beyond merely stating that a person is a collection of stored facts. For instance, knowing your date of birth does not mean you remember being born. Locke emphasized being conscious of having experienced or done something. His perspective captures a key reason why memories matter to identity: they allow us to encounter the past as our past, rather than just as facts about a person's life.

Butler questioned whether memory can truly establish the identity of the person doing the remembering.[12] Imagine I genuinely state that I remember planting a tree. Calling this a genuine memory already suggests that I was the one who planted it. If someone else planted it and I only imagined doing so, then the experience is not a true memory of my own action. Butler's point is that memory might provide evidence that I am the same person, but it cannot explain what makes me the same person if that identity is already presumed when the experience is labeled a memory.

There is also a problem with using only direct recall to assess memory changes. An experience can shape someone long after they lose the ability to remember it. A person might forget an early interaction with a teacher but still maintain interests and habits that developed from that relationship. Conversely, remembering an event does not guarantee that it continues to affect the person in the same way. Someone might clearly recall choosing a profession but no longer care about the goal that once drove that decision.

Neither example disproves everything Locke said about consciousness. They show the limitations of a simple test that only asks whether a patient still remembers an event. Such a test overlooks connections that persist after forgetting and changes that happen despite accurate recall. To account for those possibilities, any explanation must describe how memories connect with the rest of a person's mental life.

4. Parfit: Remaining Connected Without Remaining Unchanged

Parfit explains this wider concept by separating psychological connectedness from psychological continuity.[13, 14] Psychological connectedness means direct ties between a person's mental states at different points in time. For instance, recalling a conversation from the

previous day is one such tie. Carrying out an intention formed yesterday is another. Also, keeping a belief or caring about a relationship links different moments of a person's life. Memory is important, but it is just one type of connection among many.

Psychological continuity shows how these connections can last for much longer periods. Consider a person at ages ten, twenty, and forty. At twenty, they recall many experiences from childhood and still have some interests they developed at ten. At forty, they remember much of their life at twenty and maintain commitments made during those years, even if few memories from age ten remain. The forty-year-old can still have links to childhood through these intervening stages. Continuity involves a sufficiently strong, overlapping series of psychological connections; it does not require a direct memory reaching back to every earlier event.

This difference helps explain how significant changes can happen within a continuing life. A forty-year-old might have different goals than they did at ten, yet those goals emerged through years of learning, choices, and relationships. The connection is historical: the later psychological life grew from the earlier one. Mere similarity would not be enough. A stranger with the same interests and personality would not share this continuity.

Parfit also takes Butler's circularity problem seriously. His discussion of "quasi-memory" asks us to describe a recollection-like experience and its proper causal link to a past experience without first assuming both belong to the same person.[14] The goal is not to rename actual memories. Instead, it is to avoid using the conclusion (that these are the same person) as part of the evidence that leads to that conclusion. For this paper, the key point is that psychological connections must be described and examined, not simply assumed because we call an experience "my memory."

Parfit further argues that the psychological relationships important for survival do not need to treat identity as an extra, separate fact.[13, 14] This does not mean someone can be partly the same individual in the usual numerical sense. It means that direct psychological connectedness can grow stronger or weaker, and these shifts deserve attention even when there is no question about which individual continues to exist. Therefore, a patient's concern about losing an attachment can be taken seriously without claiming that treatment would literally replace the patient with another person.

This is why Parfit's ideas are useful for memory modification. Imagine someone remembers a traumatic event less clearly but still cares about the people, plans, and obligations connected to it. A change in recall would not mean all connections to that period were lost. Conversely, someone might keep a detailed memory while losing an attachment that previously made the event central to their life. Parfit's perspective can distinguish these outcomes because it considers the range of psychological connections, not just access to the remembered event.

However, describing a connection does not decide whether keeping it is good. People sometimes have reasons to update long-held beliefs or let go of commitments. Parfit's account helps identify the change; it does not make psychological continuity a requirement to stay the same. To understand why a particular connection matters to a patient, we also need to consider how they view its place in their life.

5. Ricoeur: Understanding Who I Am Through What Has Happened

Ricoeur's concept of narrative identity explores how people interpret their experiences and actions as elements of their overall life story.[15] In this context, "narrative" signifies more than just recalling events in the order they

occurred. It involves recognizing the connections between them: for example, how one experience influenced a decision, why a promise holds significance, or how a shift aligns with the kind of person someone sees themselves as. A simple list of past jobs shows what someone has accomplished. However, understanding why their desire to care for others led them to nursing reveals something about how they define themselves.

Ricoeur differentiates between sameness and selfhood. Sameness refers to characteristics that stay relatively consistent, such as familiar traits and behaviors. Selfhood allows us to grasp a person's continuity in ways that do not rely solely on unchanging qualities.[15] Keeping a promise serves as an illustration of this distinction. A person's preferences might change after making a promise, yet they may still feel responsible for upholding it. What connects them to that earlier commitment is not merely having the same desire; it is their ongoing accountability for what they agreed to do.

Narrative identity helps explain this kind of persistence through change. Consider a nurse who entered the profession after receiving care during a childhood illness. Initially, this nurse might view their career choice as repaying a debt. Later, they might come to see it as an independent commitment to helping others. The original event remains part of their life, but its significance shifts. Narrative identity allows for such reinterpretation without requiring the individual to stick to their first explanation forever.

When applied to memory modification, this framework directs attention to a question that simply retaining information cannot answer: what does this memory mean to the person now? A patient might accurately recall an accident but no longer consider it an event that must dictate every future decision. They could experience this shift as regaining possibilities that fear had previously restricted. Another person might perceive such a change as a detachment from something they still consider fundamental to their life. Narrative identity brings these differences in self-understanding into the discussion about identity.

This approach does not demand that patients create a perfectly crafted autobiography or force every experience into a single storyline. Strawson challenges both the idea that everyone experiences life as a narrative and the expectation that everyone should.[16] His point is relevant here: struggling to tell a coherent life story should not be taken as proof that someone has lost their identity. Narrative interpretation can focus on how a patient understands one relationship or commitment without requiring a story that organizes their entire life.

Nor does a satisfying story alone guarantee that all important connections have survived. A person might misunderstand why a commitment developed or offer an account that overlooks changes in their behavior. Ricoeur's perspective is more nuanced than a simple demand for a convincing story; reducing it to mere narrative coherence would misrepresent its value. Its contribution here is to explain the personal importance of continuity and change, while still allowing for examination of whether the described relationships are genuinely maintained.

6. What Each Account Alone Could Miss

Imagine a patient whose partner died in an accident. The memory of the accident triggers fear, which interferes with their daily routine. What if a therapy could lessen that fear, without preventing the patient from remembering the incident? This example explores different interpretations of psychological changes. It doesn't suggest that current treatments can consistently distinguish fear from grief or attachment.

First, we must consider a complication. The patient is both receiving treatment and

mourning, and loss itself changes people. Over time, individuals often become less involved with projects they shared with the deceased, reconsider the importance of those projects, or find their motivation for them decreases. These shifts are part of normal mourning, not memory changes. So, the example specifies that the changes described later appear soon after treatment, differ from the patient's prior grieving process, and involve the memory the treatment targeted. In reality, these conditions could only be estimated. For instance, one might note when a change began relative to treatment and compare it with the typical course of bereavement. Without such proof, an analysis could describe how the patient's connection to the past has changed, but it could not attribute that change to the treatment.

Before treatment, the patient had continued a community project started with their partner. Afterward, the patient still remembers their shared work, cares about the partner's family, and plans to finish the project. This shows significant psychological links between their past and present life. A Parfitian view confirms these connections remain. However, that finding doesn't yet answer the patient's question about whether the project still reflects who they are.

Suppose the patient once saw the project as a shared goal, but now pursues it only because quitting would feel like a betrayal of their partner. The intention to finish persists, yet its role in the patient's life has changed. Stating that the commitment continues would be accurate but incomplete. It could hide the difference between a purpose the patient embraces and a burden the patient feels unable to drop. A narrative approach highlights this difference by asking how the patient links the project to the relationship, the loss, and their future.

Now, consider the opposite scenario. After treatment, the patient says, "My partner's influence is still central to my life," and gives a familiar explanation for the project's importance. Nevertheless, they no longer make plans for the project, show interest in its progress, or keep in touch with the people involved. Assume these changes reflect a loss of attachment rather than practical obstacles, and that they meet the conditions above instead of being part of normal grief. The patient's description remains recognizable, even though some of the psychological connections it describes have weakened.

An analysis that accepted the ongoing story as enough evidence of preservation might miss this change. Parfit focuses on whether the earlier intentions and attachments actually continue, rather than just whether the patient can still describe them. The point is not that an observer should automatically dismiss the patient's account. The inconsistency itself requires explanation. The story might express a commitment the patient hopes to regain, or it might preserve an earlier self-description that no longer fits their current attachments.

These two possibilities reveal distinct errors. We could confuse existing connections with an unchanged relationship to the past. Or we could confuse an unchanged account of the past with existing connections.

7. Why the Combination Is Justified

The two perspectives appear to overlap significantly, to the point where each seems to encompass the other's domain. This part of the text will address two criticisms stemming from this perceived overlap, and then explain the practical benefits of using both together.

We begin with Parfit's theory, as it pinpoints the specific connections a treatment might change without presuming that every memory alteration automatically means a loss of identity.[13, 14]

Someone supporting Parfit's view might argue that this added element is not needed. If a

patient's understanding of the project is a belief or attitude, then Parfit's framework of psychological relations already covers it. This point accurately describes the range of the theory. Therefore, the case for narrative identity does not rest on the idea that Parfit's work excludes meaning from psychology. Instead, it argues that a description of what links stages in a life does not inherently explain how those links make that life understandable to the person living it. Narrative analysis provides this additional explanation. For example, it investigates how a patient's loyalty to their partner became tied to continuing the project, and whether they can view changing the project as consistent with that loyalty.

Similarly, a supporter of Ricoeur could argue that narrative identity already deals with actions, responsibility, and relationships with others, not just with words.[15] This objection is also valid. The situation does not suggest that Ricoeur approves of accepting any consistent story as proof of continuity. It shows why applying narrative identity requires examining the psychological connections upon which the theory relies. Parfit makes this aspect of the analysis clear: simply repeating an earlier commitment is different from actually maintaining the intention and attachment that commitment describes.

Thus, the two theories share common ground, but they perform distinct explanatory tasks. Parfit's framework allows us to differentiate between the continuation of specific psychological connections and their modification. Narrative identity clarifies the meaning of these connections within a person's understanding of their own life. The value of combining them becomes most apparent in the different questions each viewpoint prompts a clinician to ask.

A Parfitian approach focuses on connections that can be verified against the patient's actions. Which intentions formed before treatment does the patient still act on? Which attachments still manifest in their behavior, in their contact with the people involved, in their plans, and in their attention to the project's progress? And did any changes in these areas begin after treatment, or were they already occurring as part of the grieving process?

A narrative approach asks questions that cannot be answered by behavior alone. How does the patient now explain why the project matters? Does continuing it feel like an genuine expression of their values, or like an obligation they cannot escape? Could they imagine altering the project without feeling that such a change would betray their partner?

Neither set of questions, when posed in isolation, leads to the most informative query revealed by the case: where do the answers differ, and why? A patient who continues to act on a commitment they now describe as meaningless, or who calls a commitment central yet no longer acts on it, presents a contradiction that neither perspective alone would highlight. Such a contradiction does not prove that the patient has become a different person. It indicates where a clinician should look more closely: before treatment, when explaining potential changes, and afterward, in follow-up discussions that allow the patient to assess whether a change is one they welcome. These questions are not a test of identity; they are a way to make concerns about identity specific enough for discussion.

This suggests a reason to use both perspectives in this investigation, rather than proving that either philosopher's theory is generally flawed or that these are the only available methods.

This suggestion also avoids making a coherent life story a requirement for identity. A patient may clearly understand a particular attachment without organizing their entire life as a narrative, as Strawson's criticism reminds us.[16]

Similarly, a patient may have valid reasons to revise an earlier commitment. Examining both the connections and their significance clarifies what has changed; deciding whether that change is acceptable requires further ethical judgment.

8. Conclusion

The concern that modifying memories therapeutically might change a patient's identity seldom means that a different person takes their place. Instead, it involves the patient's connection to their past, which has two aspects: the ongoing intentions and attachments they hold, and the significance those connections have for them. Parfit's theory provides a foundation for exploring the first aspect, while narrative identity helps examine the second. When these two approaches are used together, their questions can reveal alterations that either method alone would overlook. This also allows patients to re-evaluate or adjust their commitments without considering such revisions a loss of self. If a specific change is acceptable remains an ethical issue, and one that current information does not yet resolve. The issue is not just whether a memory remains, but also if and how the person's connection to their past has shifted.

9. References

[1]Nader, Karim, et al. "Fear Memories Require Protein Synthesis in the Amygdala for Reconsolidation after Retrieval." Nature, vol. 406, 2000, pp. 722–726, doi.org/10.1038/35021052.

[2]Brunet, Alain, et al. "Effect of Post-Retrieval Propranolol on Psychophysiologic Responses during Subsequent Script-Driven Traumatic Imagery in Post-Traumatic Stress Disorder." Journal of Psychiatric Research, vol. 42, no. 6, 2008, pp. 503–506, doi.org/10.1016/j.jpsychires.2007.05.006.

[3]Kessler, Ronald C., et al. "Posttraumatic Stress Disorder in the National Comorbidity Survey." Archives of General Psychiatry, vol. 52, no. 12, 1995, pp. 1048–1060, doi.org/10.1001/archpsyc.1995.03950240066012.

[4]"How Common Is PTSD in Veterans?" National Center for PTSD, U.S. Department of Veterans Affairs, www.ptsd.va.gov.

[5]Koenen, Karestan C., et al. "Posttraumatic Stress Disorder in the World Mental Health Surveys." Psychological Medicine, vol. 47, no. 13, 2017, pp. 2260–2274, doi.org/10.1017/S0033291717000708.

[6]Myers, Karyn M., and Michael Davis. "Mechanisms of Fear Extinction." Molecular Psychiatry, vol. 12, 2007, pp. 120–150, doi.org/10.1038/sj.mp.4001939.

[7]Lonergan, Michelle H., et al. "Propranolol's Effects on the Consolidation and Reconsolidation of Long-Term Emotional Memory in Healthy Participants: A Meta-Analysis." Journal of Psychiatry & Neuroscience, vol. 38, no. 4, 2013, pp. 222–231, doi.org/10.1503/jpn.120111.

[8]Schiller, Daniela, et al. "Preventing the Return of Fear in Humans Using Reconsolidation Update Mechanisms." Nature, vol. 463, 2010, pp. 49–53, doi.org/10.1038/nature08637.

[9]Chalkia, Anastasia, Lukas Van Oudenhove, and Tom Beckers. "Preventing the Return of Fear in Humans Using Reconsolidation Update Mechanisms: A Verification Report of Schiller et al. (2010)." Cortex, vol. 129, 2020, pp. 510–525, doi.org/10.1016/j.cortex.2020.03.031.

[10]Kolber, Adam J. "Therapeutic Forgetting: The Legal and Ethical Implications of Memory Dampening." Vanderbilt Law Review, vol. 59, no. 5, 2006, pp. 1561–1626, scholarship.law.vanderbilt.edu/vlr/vol59/iss5/2.

[11]Locke, John. An Essay Concerning Human Understanding. 2nd ed., 1694, Book II, Ch. 27.

[12]Butler, Joseph. "Of Personal Identity." The Analogy of Religion, Natural and Revealed, to

the Constitution and Course of Nature, 1736, first dissertation.

[13]Parfit, Derek. Reasons and Persons. Oxford University Press, 1984, Part III.

[14]Parfit, Derek. "Personal Identity." The Philosophical Review, vol. 80, no. 1, 1971, pp. 3–27, doi.org/10.2307/2184309.

[15] Ricoeur, Paul. Oneself as Another. Translated by Kathleen Blamey, University of Chicago Press, 1992.

[16]Strawson, Galen. "Against Narrativity." Ratio, vol. 17, no. 4, 2004, pp. 428–452, doi.org/10.1111/j.1467-9329.2004.00264.x.

The Limits of Consent in Adaptive Deep Brain Stimulation

Jayden Lin[1*]
[1]*Bridgeland High School, Texas, United States*

Abstract

Deep brain stimulation has treated movement disorders since 1997, and in February 2025 the first adaptive system reached the United States market. Adaptive devices detect a neural biomarker and adjust stimulation on their own, with no clinician setting parameters and no action from the patient. Ethical writing on this technology has concentrated for two decades on whether stimulation changes personality, and the evidence turns out to be thin. A review of more than 1,535 articles found 64 primary studies, of which 21 reported personality change and only 8 offered significant evidence for it. This paper accepts that finding. The researchers who run adaptive stimulation trials invert the literature's priorities: asked what concerns them, 91 percent named data privacy and security and 74 percent informed consent, while personality and identity came last, at 30 percent. Four problems have gone unexamined, and three of them exist now. The first is authorship: a device acting on a patient's affect hundreds of times a day produces moods whose origins cannot be examined. The second is responsibility: when no one can trace what the device contributed to a state, no one can soundly judge the patient's capacity to decide or her culpability for what follows. The third is consent, which lapses when a trial ends or a manufacturer withdraws and leaves the patient unprotected. The fourth depends on advances not yet made. Surgical risk, not principle, confines stimulation to catastrophic illness; as that risk falls, voluntary choices will converge on socially rewarded traits and produce a narrowing that no participant selected. The paper proposes four changes: continuous consent, patient-readable device records, funded continuity of care, and indication limits fixed in advance. For each it identifies who will oppose the change and how to enforce it.

1. Introduction

In February 2025 the United States Food and Drug Administration approved the first adaptive deep brain stimulation system for Parkinson's disease [1]. The device senses local field potentials in the stimulated circuit and adjusts its output continuously, instead of delivering a fixed pattern chosen by a clinician. Medtronic, which markets the system, called the launch the largest commercial deployment of brain-computer interface technology so far, and reported that more than 40,000 patients already use the sensing-capable platform on which the adaptive system runs [1]. Whatever one makes of the framing, the approval marks the point at which closed-loop neuromodulation left the research setting for routine care.

Conventional stimulation, by contrast, has been in clinical use for decades. The FDA approved stimulation of the ventral intermediate nucleus of the thalamus for essential tremor in 1997 and extended approval to the subthalamic nucleus and globus pallidus internus for Parkinson's disease in 2002. Humanitarian device exemptions followed for dystonia in 2003 and for obsessive-compulsive disorder in 2009, the latter targeting the ventral capsule and ventral striatum [2]. Hundreds of thousands of devices have been implanted since.

What changed in 2025 was the control architecture, not the hardware in the skull. Conventional stimulation runs open loop. A clinician selects amplitude, frequency, and pulse width, the device delivers that pattern

continuously, and the patient reports back at the next appointment. Adaptive stimulation closes the loop, so that the device measures a neural signal, compares it against a threshold, and modulates output accordingly, thousands of times between clinic visits. Lozano and colleagues treat this transition as one of the field's central directions, alongside improvements in targeting and electrode design [3].

Ethical scholarship on stimulation is extensive, and almost all of it addresses one question: whether stimulation alters personality, identity, agency, authenticity, autonomy, and the self. Gilbert, Viaña, and Ineichen grouped these concepts under the acronym PIAAAS and argued that the discourse had outrun its evidence [4]. Their argument is substantially correct, and this paper accepts it. What accepting it does, though, is move the ethical problem somewhere else. The difficulties adaptive stimulation actually creates were never the difficulties that the personality debate was built to address, and settling that debate has left them largely unexamined.

Sections 3 and 4 take up the first of them, which concerns authorship. A device that regulates affect automatically produces mental states whose causal history the patient cannot reconstruct, even where she endorses each of them. Section 5 turns to responsibility, since any assessment of culpability or decisional capacity assumes that someone can trace where a person's motivational states came from. Sections 6 and 7 examine the consent instrument at both ends of its life, at implantation and at the point where a device is abandoned by its manufacturer or removed when a trial concludes. The first three problems are current and concern devices already in patients. Sections 8 and 9 turn to a fourth problem that does not yet exist: what happens once trait modulation becomes safe and widely available. Section 10 proposes a response to each problem and names who will resist it.

2. What adaptive stimulation changed

Adaptive control in psychiatry is best illustrated by a single-patient study that Scangos and colleagues published in 2021 [5]. Their patient, identified as Sarah in subsequent coverage, had severe treatment-resistant depression. The team first implanted ten temporary electrodes across candidate regions. Over the following ten days they recorded her intracranial activity while she rated her own symptoms. Gamma power in the amygdala emerged as a marker of her worst states, and stimulation delivered to the ventral capsule and ventral striatum reliably improved them. A chronic sensing and stimulation device followed, programmed to monitor the amygdala and fire a brief pulse into the striatum whenever the biomarker crossed threshold [5].

Sarah improved rapidly, and the improvement held [5]. She also acquired a device that fires hundreds of times a day. She presses nothing, decides nothing about timing, and perceives almost none of the individual events. Because the biomarker came from her own recordings, the decision rule is specific to her, fitted by a research team using cross-validated statistical models. No clinician can predict from general principles when the device will act [5]. Nothing in the argument below suggests that she should have been denied this treatment.

Under open-loop stimulation the arrangement looks quite different. A clinician chooses settings, the patient lives with the effect for days or weeks, notices it, and reports on it at her next visit. Her testimony is the mechanism through which the therapy gets tuned, which places her inside the feedback loop as a participant. Adaptive control removes that position. The loop now closes between biomarker and stimulator, and the patient becomes the substrate on which it operates.

Goering and colleagues built this into an account of relational agency, arguing that psychiatric applications are especially fraught because conscious authorship of one's own mental states matters more in psychiatry than in movement disorders [6].

Cardiac pacemakers automate a continuous decision, and nobody treats a paced heart rate as a threat to agency, so automation cannot be the whole of the concern. The difference lies in what gets regulated. A pacemaker governs a physiological variable that no one takes to be constitutive of who the patient is, while an adaptive psychiatric device governs the affective states from which intentions and actions arise.

3. The deflation of the personality-change literature

Before building on that asymmetry, the empirical record needs an accurate accounting, since the version circulating in public discussion is misleading in ways that cost patients something real.

Gilbert, Viaña, and Ineichen screened more than 1,535 articles on the putative effects of stimulation on personality and related concepts. Sixty-four qualified as first-hand primary research, meaning new clinical cases involving direct interviews, or studies involving implanted patients [4]. Commentators summarizing that dataset report that only 21 of the 64 identified personality changes after stimulation, and that only 8 of those 21 were judged to provide significant evidence [7]. The remaining 43 papers, roughly two thirds of the primary literature, failed to support a direct link [8]. Publications in theoretical neuroethics asserting such changes nonetheless increased steadily across the same period, which led the authors to conclude that the debate rested on speculative assumptions [4].

Caution is warranted, and the most frequently cited illustrative case shows why. Mantione, Figee, and Denys described a patient treated with nucleus accumbens stimulation for severe, long-standing obsessive-compulsive disorder. He developed a sudden and durable preference for the music of Johnny Cash. The preference persisted without habituation, and the authors read it as evidence for accumbens involvement in musical reward [9]. It is a genuinely interesting case, and it supports a modest claim about reward circuitry. It does not show that neurosurgeons can reshape a personality, yet it circulates in ethical discussion as though it did.

The deflation argument drew substantial replies, and the exchange clarifies what the evidence can and cannot settle. Snoek, de Haan, Schermer, and Horstkötter objected that Gilbert and colleagues had not shown the identity debate to be a bubble at all, since the 43 studies counted as negative evidence were never characterized in enough detail for readers to evaluate them [8]. Gilbert and colleagues replied that whether the evidence suffices is an epistemic question requiring replicable data. They added a point that is often overlooked: researchers who make unsubstantiated claims about personality change bear some responsibility when prospective recipients delay or refuse an effective treatment out of fear [10]. Overstating the risk carries costs, and those costs fall on patients with treatment-resistant illness who have few alternatives.

What the empirical record does contain is a set of narrower and better-supported findings. Flores Alves Dos Santos and colleagues, introducing a randomized trial of a psychoeducation programme, noted that the prior literature reports psychological and social maladjustment in roughly a third of patients after subthalamic stimulation, work disruption and marital discord among them [11].

These findings describe something real without describing surgical personality editing. Stimulation sometimes produces psychological

and social effects that patients and families find difficult, those effects vary widely across individuals, and the mechanism is often the demanding adjustment to recovery rather than any direct action of the current on character. That picture fits the evidence far better than the version reaching the public.

4. Authorship under automatic stimulation

Accepting the deflation changes which question is worth asking. If stimulation rarely rewrites a personality, then arguments premised on personality rewriting fail, and what remains is a question about authorship that applies with more force to adaptive systems than to conventional ones.

An adaptive psychiatric device detects a neural state associated with a symptom and intervenes to change it. The intervention alters, affects, shapes what a person wants and notices and decides, and actions follow from those altered states. The patient experiences the resulting actions as her own, because they do arise from her own motivational states, and the device's causal contribution is invisible from the inside. Klaming and Haselager posed the question directly for conventional stimulation, asking what implanted devices do to psychological continuity and to responsibility for action [12]. Adaptive control sharpens it, since the device now selects the timing and magnitude of each intervention on the basis of computations that neither clinician nor patient observes as they occur.

What she cannot do is locate the device's contribution to any particular state she finds herself in. She knows it is implanted and knows in general terms what it does, so concealment is not the issue. Between clinic visits the device makes a large number of decisions that no human being reviews, each individually small. Their cumulative effect is a maintained affective baseline, produced jointly by her own neurophysiology and by an algorithm she did not write and cannot inspect.

Autonomy has at least two components, and adaptive stimulation pulls them apart. One is endorsement, meaning that a person affirms the state she is in and would choose it if asked. The other is authorship, meaning that the state arose through processes she can in principle account for. Adaptive stimulation can preserve endorsement completely while degrading authorship, and a patient may sincerely endorse every mood the device maintains while having no account whatever of why she is in it. One might object that this proves too much, since every mental state has causes the person did not author, among them sleep, medication, and upbringing, and nobody treats an antidepressant as a threat to authorship. What distinguishes the implant is not the presence of an unchosen cause but the patient's inability to interrogate it. She can stop a drug tomorrow, its mechanism is public, and its population-level effects are documented. An adaptive device runs on rules set by a manufacturer, adjustable remotely by a clinician, often proprietary, and unavailable to her in any readable form.

5. Responsibility for device-mediated action

Separating endorsement from authorship has consequences beyond the philosophical, because several practices that matter a great deal to patients depend on being able to reason about where a person's mental states came from.

Criminal and civil responsibility make the clearest case. Klaming and Haselager concluded that stimulation generally does not undermine responsibility for action or mental competence, since patients typically retain the capacities that responsibility ascription requires, though they flagged edge cases in which stimulation-induced impulsivity complicates both competence assessment and culpability [12]. They reached that conclusion about open-loop devices, where the stimulation

parameters are fixed, documented, and chosen by an identifiable clinician who can be asked to explain them. Adaptive systems weaken each of those conditions. Parameters change continuously, a classifier rather than a clinical judgment drives the change, and the rationale for any given intervention exists only as a threshold crossing in a device log. That log is typically inaccessible to the patient and often proprietary to the manufacturer.

Writing about learning automata, Matthias described situations in which the operator of a system cannot reasonably be held responsible for behaviour she could neither predict nor control, while no other party is obviously better placed to bear that responsibility instead [13]. Something similar arises here. The analogy should not be pushed too far. A patient with an adaptive implant is no mere operator of an external machine, and the states the device modulates are her own states, not the outputs of a separate agent. The practical point is narrower. When a forensic psychiatrist, a treating clinician, or a court tries to reconstruct why a patient acted as she did, part of the relevant causal history now sits in a record that none of them can read.

Existing legal doctrine gives implanted patients very little to work with. Rotenberg and colleagues, writing in Science, observe that the Supreme Court's decision in Riegel v. Medtronic sharply limits product-liability claims against devices approved through the premarket approval pathway. Most high-risk neural devices follow that pathway. They propose a no-fault compensation scheme for patients injured by implantable brain-computer interfaces, on the grounds that existing tort mechanisms will not reach them [14, 15]. Their argument concerns physical injury, not device-mediated action, but the structural observation carries over. Where harm emerges from the interaction of a patient, a clinician, and an adaptive algorithm, liability tends to distribute in a way that leaves the patient holding the residue.

The argument is limited, since no widely reported criminal case has turned on stimulation-induced action as a defence. Nothing in the current evidence suggests that patients with implants are routinely absolved of responsibility, or that they should be. The claim here is narrower: the epistemic basis on which responsibility and capacity get assessed has already degraded for every patient whose device runs a closed loop. Courts have not yet faced this gap, but clinicians face it whenever they assess one of these patients' capacity to decide.

6. The structure of consent for a continuously acting device

Informed consent in medicine authorizes a particular intervention. The patient learns what will be done, what risks attach, and what alternatives exist, then agrees or declines, on the assumption that what she is authorizing is a bounded procedure with a beginning and an end. Adaptive stimulation has no such shape, and the mismatch is more than formal.

Muñoz and colleagues interviewed twenty-three researchers running adaptive stimulation trials and catalogued their ethical concerns, with results that are instructive for anyone deciding which problems in this area require attention [16]. Data privacy and security led the list at 91 percent, followed by risks and safety at 83 percent and informed consent and adequate patient understanding at 74 percent. Automaticity and device programming appeared in 65 percent of the interviews, patient autonomy and control over stimulation in 57 percent, and post-trial access and device maintenance in 39 percent. Effects on personality and identity, the topic that has dominated theoretical literature for two decades, appeared in only 30 percent [16]. The people building these systems worry most about the structural problems this paper identifies, and

least about the one the ethics literature has emphasized.

Those same researchers raised the difficulty of ensuring that patients consent appropriately to continuous alterations in stimulation, and the authors listed enhanced ongoing consent among four priority responses [16]. The recommendation has not been implemented in any binding form. A patient implanted with a commercially approved adaptive system signs a consent document before surgery, nothing requires that consent to be revisited as the device operates, and nothing gives her any right of access to a record of what it has done.

Manufacturers establish threshold settings and classification rules, clinicians adjust them, sometimes remotely, and no requirement exists that a patient be able to read a log of when stimulation fired or what triggered it. Nor can she deliberate usefully about a policy she cannot observe in operation, since consent to automatic stimulation in general is not consent to any particular intervention, and a patient who cannot discover which interventions occurred has no material on which to reason about whether she wants them to continue. Time compounds both problems. Her authorization was momentary while the relationship it created extends indefinitely, a mismatch that becomes acute at the far end of the relationship in ways the next section examines.

7. Abandonment and the far end of the consent relationship

If consent at implantation authorizes an ongoing relationship, the obligations that relationship generates should extend as long as it does. In practice they do not, and the resulting failures are documented well enough to count as a known feature of the field rather than a hypothetical one.

Working from a systematic review and a consensus process convened at a Royal Society summit on neural interfaces, Okun and colleagues developed an operational definition of implanted neurological device abandonment. It spans deep brain, vagal nerve, and spinal cord stimulation. Abandonment following a failed research trial, they note, has to be defined contingent on the circumstances of that trial [17]. That such a definition was needed at all indicates how routine the underlying problem has become.

Second Sight Medical Products implanted roughly 350 patients with its Argus II retinal prosthesis, discontinued the product in 2019, and came close to collapse in 2020, leaving recipients with implanted hardware that could not be repaired, upgraded, or supported [18]. Nuvectra Corporation, whose Algovita spinal cord stimulation system had been implanted in at least three thousand patients, filed for bankruptcy in November 2019. Patients whose batteries failed or who needed reprogramming were stranded until another firm acquired the assets [19]. A third case involves publicly funded research instead of commercial failure. Ian Burkhart, the first person with paralysis to regain volitional hand movement through a cortically controlled neural bypass, received his implant in 2014 as part of a clinical trial. It was explanted in 2021 after roughly seven years, when the funding ended, and he has described the loss as a capacity removed rather than a device returned [20].

These failures have generated a body of scholarship. Lázaro-Muñoz and colleagues argued that no guidelines ensure participants continued access to implants that benefited them, and called for shared responsibility among sponsors, investigators, and manufacturers [21]. Gilbert has argued that forced explantation can itself produce self-estrangement, since a patient who has integrated a device into her sense of agency loses more than a piece of hardware when it comes out [22].

Every element of this intensifies under adaptive control. A patient whose affective baseline is maintained by a closed loop depends on it more heavily than one receiving fixed stimulation, since withdrawal removes an ongoing regulatory process instead of a setting she has adjusted to. The algorithms that make adaptive devices valuable are also the components most likely to be proprietary, which makes third-party maintenance harder once a manufacturer exits. Consent obtained at implantation says nothing about any of this, and the patient who signed it has no way of discovering what will happen to her if the company that built her device stops answering the phone.

8. Indication drift and the boundary between treatment and enhancement

The first three problems concern patients who already have these devices. The fourth concerns who will receive them, and it arises only when two conditions are met: surgical risk low enough to justify implanting people who are not severely ill, and targeting precise enough to change one trait without disturbing the rest.

Stimulation is currently reserved for severe, treatment-resistant illness, a restriction often described as though it followed from some principle about the proper aims of medicine. What actually produces it is the risk-benefit calculation of intracranial surgery. Researchers interviewed by Kostick-Quenet and colleagues said as much unprompted. One observed that the risk-benefit ratio for a person already disabled by a pathological disorder differs entirely from the ratio for a healthy person. Another noted that the invasiveness of the procedure deters casual uptake [23]. A threshold that exists because of a cost will move when that cost falls.

The range of conditions under investigation has already moved well beyond movement disorders, and now takes in treatment-resistant depression, obsessive-compulsive disorder, Tourette syndrome, substance use disorder, anorexia nervosa, schizophrenia, and refractory aggression [24]. Schlaepfer and colleagues showed that stimulation of reward circuitry alleviates anhedonia in refractory depression [25]. The TRANSCEND trial for treatment-resistant depression implanted its first patient at Mount Sinai in early 2025 and now recruits across twenty-five sites in the United States [26].

Efficacy within existing indications is moving too, which strengthens the case for intervening earlier and in less severe presentations. A recent meta-analysis across neurological and psychiatric indications found the largest effect sizes in essential tremor among neurological conditions and in Tourette syndrome among psychiatric ones. Its authors concluded that wider clinical use in obsessive-compulsive disorder and Tourette syndrome is justifiable on current evidence [27].

Stimulation of reward circuitry for anhedonia modulates the capacity to experience pleasure [25]. Stimulation for obsessive-compulsive disorder modulates how intensely a person responds to perceived contamination or wrongness [9]. Stimulation for refractory aggression modulates the threshold at which anger produces action [24]. Each gets described as symptom control, and in the current patient population each description is accurate, because the symptoms in question are catastrophic. The underlying mechanism, however, does not restrict the effect to catastrophic cases. The same dial that lifts anhedonia to normal can lift normal to elevated, which is why Synofzik, Schlaepfer, and Fins asked how happy is too happy in the context of accumbens stimulation more than a decade ago [28].

Parens framed enhancement as intervention that improves a capacity beyond

what health requires [29]. The formulation is standard, and it proves unstable in practice, since health gets defined by the same clinical community that decides which conditions warrant intervention. No committee will announce a transition from treatment to enhancement. Approvals will extend one indication at a time, each defensible on its own evidence, and the severity threshold will drift downward as a byproduct.

Researchers in the field expect this drift. Kostick-Quenet and colleagues interviewed twenty-three adaptive stimulation researchers and found that 70 percent believed enhancement applications will eventually be feasible, several noting that development for non-clinical purposes is already underway in military contexts [23]. Their ethical concerns clustered around safety and security, the sense that enhancement is unnecessary or aberrant, and fairness and distributive justice. All twenty-three held that stimulation should not be used for enhancement until researchers understand brain target localization and function considerably better than they do now [23]. Their unanimity says something about present capability, not about principle, and it is exactly the kind of restriction that dissolves as capability improves.

9. Convergence without coercion

Suppose the drift completes, and stimulation becomes precise enough and safe enough to modulate traits in people who are not catastrophically ill. The standard objections at that point concern access and inequality. One researcher interviewed by Kostick-Quenet and colleagues worried about a future in which a person effectively needs a device to succeed as a college student. Others expected the technology to amplify existing class divisions [23]. Those concerns are serious, though they are not the deepest problem, since equal access would answer them. The deeper problem arises precisely when access is equal.

People modify toward what their environment rewards, and the rewarded traits are well known: warmth, confidence, low neuroticism, high agreeableness, sustained focus. Kramer identified this dynamic in psychopharmacology three decades ago. He coined the term cosmetic psychopharmacology for the use of drugs to move a person from one normal state to another that is more socially favoured. He also observed how well antidepressant effects such as increased assertiveness fit the demands his culture placed on women [30]. Kramer has since acknowledged that pharmacology never delivered the capacity to sculpt personality to order, which identifies the limitation as technical, and technical limitations are the kind that engineering eventually removes.

One researcher in the Kostick-Quenet interviews reached the aggregate concern without being asked about it. Treating depression and anxiety is already a form of mood enhancement, he noted, considered reasonable because it addresses symptoms. He then asked whether it would be better if everybody were a little happier, and whether society itself changes at that level [23]. The answer requires looking at the structure of the harm rather than at any individual case.

Every individual modification in the scenario is voluntary, informed, and endorsed by the person afterward. Nobody is coerced, no participant is made worse off by her own lights, and each can articulate a good reason for her choice. The choices are not independent of one another, though, because they respond to a shared standard of what makes a person employable, likeable, and successful. As more people meet that standard through modification, the standard itself shifts, since the reference distribution against which normality gets judged is now composed of modified people. Pressure to modify then increases for those who have not. What results is a collective action problem in

which every individual step is rational and consensual while the aggregate outcome is a narrowing that no participant selected. A natural response is to ask who has actually been harmed, since each person chose her own modification and endorses the result; a preference for the current distribution of human personality can start to look like nostalgia dressed as ethics. Most of that should be granted. Individual modification is not the target here, and nothing below argues for withholding relief from any particular person. The harm sits in the recursion rather than in any individual choice, in a norm that regenerates from its own output and therefore has no floor and no stable reference point, so the pressure it exerts increases instead of settling at an equilibrium. What follows from this is not a case for restricting individual choice but for fixing the range of permissible indications through a mechanism that individual choice cannot ratchet upward, a matter of institutional design rather than a judgment about the value of any particular personality.

Traits that societies penalize are not distributed randomly across moral categories, and several of them do considerable moral work. Disagreeableness sustains dissent. Anxiety sustains vigilance about risk. Scrupulosity in non-pathological amounts sustains the refusal to go along with what one takes to be wrong. A system that lets individuals opt out of socially costly traits, one person at a time, would remove the substrate of moral disagreement without anyone having decided to remove it. The argument needs only the observation that preferences are endogenous to the distribution from which they sample.

10. Governance implications

The analysis points toward four changes in how these devices are governed, each addressing a failure identified above. The first three address harms that exist now, and regulators can impose most of them with authority they already hold. The fourth guards against a harm that does not exist yet. Each change runs against someone's interest, so each has stayed unimplemented; the paragraphs below name that interest and how to overcome it.

Consent for adaptive systems should be continuous and revisable. Muñoz and colleagues recommended enhanced ongoing consent on the basis of researcher interviews [16], and it should become a condition of approval. Trial sponsors will resist. Adaptive psychiatric studies are small, sometimes a single patient, as in the report by Scangos and colleagues [5], and teams keep adjusting their detection thresholds and stimulation rules as data accumulate. Every re-consent gives a participant a chance to withdraw, and in a study of a few people one withdrawal can cost the study its statistical power. A narrower requirement lowers that cost. The patient authorizes a range of thresholds and targets at surgery, and only a change outside the range triggers re-consent. FDA has imposed conditions like this before: in 2021 it restricted the sale of breast implants to providers who review a decision checklist with each patient before surgery [31]. The same power could limit adaptive devices to centres that record re-consent whenever a configuration leaves its agreed range.

Patients should have read access to their device records, including when stimulation fired and which signal set it off. This answers the authorship problem of Section 4 and the responsibility problem of Section 5 at once. A patient who can see what her device contributed to a given state keeps authorship of it even under automatic control, and a clinician or court reconstructing her motivation has a record to work from. Manufacturers will resist, as they did in cardiology. Hugo Campos, implanted with a Medtronic defibrillator, spent years unable to see the data it recorded about his own heart. He got his raw data in 2013, and only after confronting

Medtronic's chief executive in public on Twitter [32]. Two incentives keep device logs closed. Aggregated across thousands of patients, the logs could serve as training data for the next generation of detection algorithms, which makes them commercially valuable. They are also evidence: the product-liability suits that survive Riegel are the ones a readable log would support [15]. FDA can override both incentives. The catch-all in its post-approval regulations lets it impose any requirement necessary for a device's continued safety and effectiveness, and a log that lets patients and clinicians catch a malfunction qualifies. A narrow log exposes nothing proprietary: the patient needs the time of each event, the biomarker value that crossed threshold, and the dose delivered, none of which reveals the classifier.

Continuity-of-care obligations should attach to the implant, not to a trial's enrollment window or a product's commercial life. Okun and colleagues have defined device abandonment [17], and Lázaro-Muñoz and colleagues have argued for shared responsibility among everyone who benefits from an implant [21]. That means funded maintenance and explantation plans, established before the first patient is enrolled. The resistance is financial. Money reserved for battery changes, reprogramming, and eventual removal cannot go into development, and the firms least able to reserve it are the ones most likely to fail, as Nuvectra did [19]. Other industries face the same incentive and require money up front. A surface coal mine cannot get its permit until the operator posts a bond covering reclamation [33], because an operator facing an expensive cleanup has every reason to disappear before paying for it. An implant sponsor could post a surety bond or carry an insurance policy instead of locking up cash, so a small firm pays a premium and that premium prices its own risk of failure. NIH can require this of the trials it funds now; binding every commercial sponsor would take legislation, as reclamation bonding did for mining.

Indication limits should be set at the population level and fixed before precision improves. This is the least conventional proposal and the most important one. Once trait modulation is low-risk, every case-by-case argument will favour extension, because each person asking for it presents a sympathetic case and no identifiable person is harmed by saying yes. Only a threshold fixed in advance survives that pressure, and it draws the widest opposition of the four. Manufacturers gain a market with each new indication, as stimulation for obsessive-compulsive disorder shows. Medtronic reached the market in 2009 through the humanitarian device exemption, a pathway meant for rare conditions that asks only for probable benefit and waives proof of effectiveness, and Fins and colleagues argued the route let the company skip the trials a full approval would have demanded [34]. Physicians will resist too, because the law lets them use an approved device off-label. Coverage reaches both. Medicare already enforces a severity floor for stimulation in Parkinson's disease, paying only for advanced disease with disabling symptoms that persist despite optimal medication [35]. Applied to psychiatric indications, with the floor written in numerical severity scores and changeable only through public notice and comment, the same mechanism would enforce the fixed threshold. None of this stops a wealthy patient from travelling to a more permissive country. It stops the approval system from lowering the threshold on its own.

11. Conclusion

The neuroethics literature on deep brain stimulation asked whether the technology changes who patients are, and the empirical answer is that it rarely does so in the dramatic sense the question implies [4, 7]. Accepting that

answer does not close the ethical inquiry so much as redirect it toward problems the personality question obscured.

Three of those problems exist now. Adaptive devices make continuous, unreviewed decisions about a patient's affective states, while the consent she signed was written for a discrete procedure [16]. The same opacity that costs her the authorship of her own moods erodes the basis on which clinicians and courts judge her responsibility for what those moods lead her to do [12, 14]. At the far end of the same relationship, patients who depend on an implant can be left without support when a trial closes or a manufacturer fails [17, 21]. The fourth problem has yet to emerge. Surgical risk, not principle, is what confines stimulation to catastrophic illness, and that threshold will move as risk falls, one defensible approval at a time. Should trait-level modulation become available, individually voluntary choices will converge on socially rewarded traits, and the aggregate outcome will be a narrowing that no participant chose and no institution authorized.

The technology is neither as dangerous nor as safe as it is usually made out to be. It works, it helps people with severe illness, and the case reports circulating in ethical discussion overstate its power to remake a person [9, 10]. The problems worth addressing are the ones that arrive without anyone having decided on them.

12. Discussion

The argument in Section 9 is anticipatory, and anticipatory ethics carries a known failure mode. Wexler's caution against analyzing capabilities that do not exist applies directly to any claim about trait modulation in healthy populations [36]. The researchers interviewed by Kostick-Quenet and colleagues were unanimous, too, that current understanding of brain localization is inadequate for such applications [23]. Should precision plateau, the population-level argument will simply continue to be a predicted outcome.

The empirical basis for the deflation argument also has limits that its critics identified accurately. Snoek and colleagues noted that the primary studies counted as failing to support a link were never characterized in sufficient detail for readers to assess them [8]. Identity changes may also be hard to detect with the instruments currently available. A conclusion that the evidence is thin is compatible with a conclusion that the measurement is inadequate. Both may be true at once, and the appropriate response is better instrumentation, not alarm or dismissal [7].

Prospective studies with adequate control groups and validated personality instruments would clarify what adaptive stimulation does to affect over time, which the current literature cannot establish. Empirical work on patient understanding of adaptive systems would test whether the consent gap identified in Section 6 corresponds to an actual deficit in comprehension. Comparative analysis of how different jurisdictions handle indication expansion would indicate whether a threshold fixed in advance is achievable in practice or merely desirable in principle.

The argument generalizes beyond deep brain stimulation, since any technology capable of modulating socially valued traits under individual control faces the same collective action structure, and any implanted device creating dependence faces the same abandonment problem. Stimulation is the instructive case because it is invasive enough to have produced an explicit severity threshold, which makes the erosion of that threshold visible in a way it is not for pharmacology or for noninvasive stimulation.

13. References

[1] “Medtronic Earns U.S. FDA Approval for the World’s First Adaptive Deep Brain

Stimulation System for People with Parkinson's." Medtronic News, 2025, https://news.medtronic.com/2025-02-24-Medtronic-earns-U-S-FDA-approval-for-the-worlds-first-Adaptive-deep-brain-stimulation-system-for-people-with-Parkinsons.
[2] "Deep Brain Stimulation (DBS) for the Treatment of Parkinson's Disease and Other Movement Disorders | National Institute of Neurological Disorders and Stroke." Www.Ninds.Nih.Gov, https://www.ninds.nih.gov/about-ninds/what-we-do/impact/ninds-contributions-approved-therapies/deep-brain-stimulation-dbs-treatment-parkinsons-disease-and-other-movement-disorders.
[3] Lozano, Andres M., et al. "Deep Brain Stimulation: Current Challenges and Future Directions." Nature Reviews Neurology, vol. 15, no. 3, Jan. 2019, pp. 148–60, https://doi.org/10.1038/s41582-018-0128-2.
[4] Gilbert, Frederic, et al. "Deflating the 'DBS Causes Personality Changes' Bubble." Neuroethics, vol. 14, no. S1, Springer Science and Business Media LLC, June 2018, pp. 1–17, https://doi.org/10.1007/s12152-018-9373-8.
[5] Scangos, Katherine W., et al. "Closed-Loop Neuromodulation in an Individual with Treatment-Resistant Depression." Nature Medicine, vol. 27, Oct. 2021, pp. 1–5, https://doi.org/10.1038/s41591-021-01480-w.
[6] Goering, Sara, et al. "Staying in the Loop: Relational Agency and Identity in Next-Generation DBS for Psychiatry." AJOB Neuroscience, vol. 8, no. 2, Informa UK Limited, Apr. 2017, pp. 59–70, https://doi.org/10.1080/21507740.2017.1320320.
[7] Kubu, Cynthia S., et al. "Pragmatism and the Importance of Interdisciplinary Teams in Investigating Personality Changes Following DBS." Neuroethics, vol. 14, no. 1, 2021, pp. 95-105, doi:10.1007/s12152-019-09418-3.
[8] Snoek, Anke, et al. "On the Significance of the Identity Debate in DBS and the Need of an Inclusive Research Agenda. A Reply to Gilbert, Viana, and Ineichen." Neuroethics, May 2019, https://doi.org/10.1007/s12152-019-09411-w.
[9] Mantione, Mariska, et al. "A Case of Musical Preference for Johnny Cash Following Deep Brain Stimulation of the Nucleus Accumbens." Frontiers in Behavioral Neuroscience, vol. 8, May 2014, https://doi.org/10.3389/fnbeh.2014.00152.
[10] Gilbert, Frederic, et al. "Deflating the Deep Brain Stimulation Causes Personality Changes Bubble: The Authors Reply." Neuroethics, vol. 14, no. S1, Springer Science and Business Media LLC, June 2020, pp. 125–36, https://doi.org/10.1007/s12152-020-09437-5.
[11] Santos, Juan E, et al. "Tackling Psychosocial Maladjustment in Parkinson's Disease Patients Following Subthalamic Deep-Brain Stimulation: A Randomised Clinical Trial." PLOS ONE, vol. 12, no. 4, Public Library of Science, Apr. 2017, pp. e0174512–12, https://doi.org/10.1371/journal.pone.0174512.
[12] Klaming, Laura, and Pim Haselager. "Did My Brain Implant Make Me Do It? Questions Raised by DBS Regarding Psychological Continuity, Responsibility for Action and Mental Competence." Neuroethics, vol. 6, no. 3, Sept. 2010, pp. 527–39, https://doi.org/10.1007/s12152-010-9093-1.
[13] Matthias, Andreas. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata." Ethics and Information Technology, vol. 6, no. 3, Sept. 2004, pp. 175–83, https://doi.org/10.1007/s10676-004-3422-1.
[14] Rotenberg, A., et al. "A Liability Framework for High-Risk Neural Devices." Science, vol. 388, no. 6752, American Association for the Advancement of Science (AAAS), June 2025, pp. 1136–38, https://doi.org/10.1126/science.adt8110.
[15] "U.S. Reports: Riegel v. Medtronic, Inc., 552 U.S. 312 (2008).." The Library of Congress, 2015, https://www.loc.gov/item/usrep552312/.

[16] Muñoz, Katrina A., et al. "Researcher Perspectives on Ethical Considerations in Adaptive Deep Brain Stimulation Trials." Frontiers in Human Neuroscience, vol. 14, Nov. 2020, https://doi.org/10.3389/fnhum.2020.578695.
[17] Okun, Michael S, et al. "Definition of Implanted Neurological Device Abandonment." JAMA Network Open, vol. 7, no. 4, American Medical Association, Apr. 2024, pp. e248654–54, https://doi.org/10.1001/jamanetworkopen.2024.8654.
[18] Strickland, Eliza, and Mark Harris. "Their Bionic Eyes Are Now Obsolete and Unsupported." IEEE Spectrum, 15 Feb. 2022, https://spectrum.ieee.org/bionic-eye-obsolete.
[19] Drew, Liam. "Abandoned: The Human Cost of Neurotechnology Failure." Www.Nature.Com, 6 Dec. 2022, https://www.nature.com/immersive/d41586-022-03810-5/index.html.
[20] "How It Feels to Have a Life-Changing Brain Implant Removed." MIT Technology Review, https://www.technologyreview.com/2023/05/26/1073655/how-it-feels-to-have-a-life-changing-brain-implant-removed/.
[21] Lázaro-Muñoz, Gabriel, et al. "Continued Access to Investigational Brain Implants." Nature Reviews Neuroscience, vol. 19, no. 6, Apr. 2018, pp. 317–18, https://doi.org/10.1038/s41583-018-0004-5.
[22] Gilbert, Frederic. "Self-Estrangement & Deep Brain Stimulation: Ethical Issues Related to Forced Explantation." Neuroethics, vol. 8, no. 2, Oct. 2014, pp. 107–14, https://doi.org/10.1007/s12152-014-9224-1..
[23] Kostick-Quenet, Kristin, et al. "Researchers' Ethical Concerns About Using Adaptive Deep Brain Stimulation for Enhancement." Frontiers in Human Neuroscience, vol. 16, Apr. 2022, https://doi.org/10.3389/fnhum.2022.813922.
[24] Mahoney, James J., et al. "Deep Brain Stimulation for Psychiatric Disorders and Behavioral/Cognitive-Related Indications: Review of the Literature and Implications for Treatment." Journal of the Neurological Sciences, Apr. 2022, p. 120253, https://doi.org/10.1016/j.jns.2022.120253.
[25] Schlaepfer, Thomas E, et al. "Deep Brain Stimulation to Reward Circuitry Alleviates Anhedonia in Refractory Major Depression." Neuropsychopharmacology, vol. 33, no. 2, Apr. 2007, pp. 368–77, https://doi.org/10.1038/sj.npp.1301408.
[26] Brzostowicki, David. "Can Deep Brain Stimulation Finally Deliver for Depression?" Medscape, 19 Aug. 2026, https://www.medscape.com/viewarticle/can-deep-brain-stimulation-finally-deliver-depression-2026a1000ssk.
[27] Froschauer, Victoria, et al. "Clinical Efficacy of Deep Brain Stimulation for Psychiatric and Neurological Indications on a Meta-Analytic Level." International Journal of Psychiatry in Clinical Practice, Informa UK Limited, Aug. 2026, pp. 1–14, https://doi.org/10.1080/13651501.2026.2710644.
[28] Synofzik, Matthis, et al. "How Happy Is Too Happy? Euphoria, Neuroethics, and Deep Brain Stimulation of the Nucleus Accumbens." AJOB Neuroscience, vol. 3, no. 1, Jan. 2012, pp. 30–36, https://doi.org/10.1080/21507740.2011.635633.
[29] Parens, E. "Is Better Always Good? The Enhancement Project." The Hastings Center Report, vol. 28, no. 1, 1998, pp. S1–17, https://pubmed.ncbi.nlm.nih.gov/9539044/. PubMed.
[30] Kramer, Peter D. Listening to Prozac: A Psychiatrist Explores Antidepressant Drugs and the Remaking of the Self. Viking Adult, 1993.
[31] Center for Devices and Radiological Health. "Things to Consider Before Getting Breast Implants." U.S. Food and Drug

Administration, FDA, 21 Feb. 2024, https://www.fda.gov/medical-devices/breast-implants/things-consider-getting-breast-implants.
[32] Parmar, Arundhati. "How One Tweet from Medtronic CEO Omar Ishrak Changed A Patient's World." Medical Device and Diagnostic Industry, 11 July 2013, https://www.mddionline.com/cardiovascular/how-one-tweet-from-medtronic-ceo-omar-ishrak-changed-a-patient-s-world.
[33] Office of the Law Revision Counsel. "30 U.S. Code § 1259, Performance Bonds." LII / Legal Information Institute, https://www.law.cornell.edu/uscode/text/30/1259.
[34] Fins, Joseph J., et al. "Misuse of the FDA's Humanitarian Device Exemption in Deep Brain Stimulation for Obsessive-Compulsive Disorder." Health Affairs (Project Hope), vol. 30, no. 2, Feb. 2011, pp. 302–11, https://doi.org/10.1377/hlthaff.2010.0157.
[35] "NCD - Deep Brain Stimulation for Essential Tremor and Parkinson's Disease (160.24)." Cms.Gov, 2019, https://www.cms.gov/medicare-coverage-database/view/ncd.aspx?NCDId=279.
[36] Wexler, Anna. "Separating Neuroethics from Neurohype." Nature Biotechnology, vol. 37, no. 9, Nature Portfolio, Aug. 2019, pp. 988–90, https://doi.org/10.1038/s41587-019-0230-z.

"Research Use Only:" Mislabelling in Counterfeit Peptides in the Grey Market

Audrey Tan[1]
[1]*Madison High School, Vienna, VA 22181*

Abstract

Therapeutic peptides are a unique class of pharmaceutical drugs made of amino acids, forming parts of proteins that carry out biological functions. First discovered in 1921 with insulin, research on peptides has continued well through the present day, where they serve a wide range of applications spanning diabetes, cardiovascular diseases, and more. However, they have also become popular within the "grey market," where unapproved and unregulated peptides are commonly sold through online vendors and promoted in online biohacking and wellness spaces. This article discusses the ethics and history behind the illegal sale of peptides, and highlights the effectiveness of a recent recommendation by the Food and Drug Administration's (FDA) Pharmacy Compounding Advisory Committee (PCAC) to move six new peptides from Category 2 (prohibited for compounding) to Category 1 (compounding permitted by pharmacies) on slowing the "grey market."

1. Introduction

Peptides have long played an important role in medicine since the discovery and development of insulin therapy in 1921.[1] Since then, the category of peptide therapies have expanded to include over 80 peptide drugs approved worldwide, including synthetic peptides in addition to natural ones, with groups ranging from GLP-1 receptors to MC4R receptors. Similarly, their uses span a large number of applications, like diabetes, migraines, various types of cancer, and more, as they commonly serve as hormones, growth factors, neurotransmitters, ion channel ligands, or anti-infective agents.[1]

Peptide development continues to evolve as more of them are developed. They initially mimicked hormones already present in nature, but major breakthroughs in molecular biology, peptide chemistry and more technologies now allow research to intentionally design and develop novel peptides with desirable biochemical and physiological properties.[1]

However, as peptide development has grown, so has unregulated peptide use. Unapproved and unregulated peptides are increasingly sold through the "grey market" and promoted online in biohacking and wellness spaces; despite a lack of supporting evidence, they claim to support fat loss, cognition, longevity, and more,[2] raising several ethical and safety concerns about these unregulated drugs.

In the United States, peptides for human consumption are managed by the FDA, who have attempted to combat the peptide "grey market" through both suppression, enforcement against illegal products, and legitimization, efforts to establish regulated pathways for several of these illegal peptides.[2] Both were used in the past year. For example, the FDA has issued warning letters against companies, like Gram Peptide, who were selling unapproved peptides,[3] and the FDA's Pharmacy Compounding Advisory Committee (PCAC) voted in July 2026 to recommend several peptides be moved from Category 2, where compounding is prohibited, to Category 1. The

committee voted on seven peptides, BPC-157, MOTS-c, KPV, TB-500, Epitalon, Semax, and Emideltide, and Emideltide was the only peptide not recommended by the committee.[4, 5]

2. The Current State of Peptides

2.1 Uses of Peptides

Therapeutic peptides are a unique class of pharmaceutical drugs made of amino acids, forming parts of proteins that carry out biological functions.[6] They can act as hormones, growth factors, neurotransmitters, and more, serving a wide range of applications as they're used to both treat and diagnose conditions like diabetes and cardiovascular diseases.[1] In addition to their versatility, peptides are also valued for their small size, immunogenicity, and comparatively lower production costs (in comparison to biologics, which perform in a similar mechanism to peptides); however, main drawbacks of these peptides include weak membrane permeability and poor in vivo stability.[1] Thus, much of peptide research works to combat peptides' low stability,[7] as though the field dates back to 1921, when Frederick Banting first isolated insulin,[1] peptide research continues to grow.

In addition to their rising popularity in medicine and pharmacy, peptides have also gained popularity in online, illegal spaces, where the unregulated drugs are often promoted in biohacking, self-medication, and other online wellness spaces.[2] The seven peptides reviewed by the FDA's PCAC in July, BPC-157, MOTS-c, KPV, TB-500, Epitalon, Semax, and Emideltide, though not yet approved by the FDA, are already sold and distributed through these online spaces often called the "grey market."[8] They claim to promote a myriad of benefits, like tissue repair, cognitive effects, metabolism, slowed aging, and more, though evidence is limited to only a few human trials, at best.[9, 10, 11]

2.2 Current Policies

The sale and distribution of peptides is regulated by the FDA under Section 503A of the federal Food, Drug, and Cosmetic Act (FDCA), which established legal protections and increased federal regulation on compounding pharmacies by defining two categories: 503A and 503B. The former is more common and includes traditional compounding pharmacies, while the latter includes "outsourcing facilities." Since 503A pharmacies, unlike 503B pharmacies, cannot compound in bulk, they are subject to fewer regulations and do not have to follow Current Good Manufacturing Practices (cGMP); thus, 503A pharmacies often face greater inconsistency in key safety areas like facility design, disinfecting compounding areas, and personnel cleaning and garbing.[32, 33] In addition, unlike retail pharmacies, which sell standardized brand-name or generic drugs, 503A pharmacies compound customized products using drugs with an applicable USP or NF monograph, components of already-FDA-approved drugs, or drugs on the 503A Bulks List.[30] The drugs pending addition to this Bulks List are grouped into three categories: category 1, 2, and 3. Peptides in category 1 are permitted for compounding, those in category 2 are prohibited, and those in category 3 are under review.[12] Additionally, ingredients without a monograph nor within an approved drug or the three categories, like AOD-9604, an anti-obesity peptide, are illegal to be compounded but nevertheless commonly sold.[34] The seven peptides listed above and reviewed by the committee were once in category 2 before their nomination was withdrawn, pausing the FDA evaluation process.[13]

2.3 Illegal Peptide Distribution, The "Research-Use-Only" Loophole and Misleading Labels

Though prohibited by the FDA, these category 2 peptides are still distributed across the U.S. through the "grey market," where many peptide products can be sold online through direct-to-consumer websites, research-chemical vendors, social media-linked sellers, and more, often with misleading labels like "Research-Use-Only" in an attempt to avoid legal repercussions and reassure consumers.[2] These products are unregulated, often lacking the necessary regulatory review, manufacturing oversight, or postmarketing surveillance required for approved drugs, but still advertise through unsupported certificates of analysis, purity claims, and biomedical language.[2]

However, the "for research purposes only" and "not for human consumption" labels found across many of these "grey market" peptide vendors[14] provides no legal protection, as the "Research-Use-Only" label does not apply as a legal exemption if the product's intended use is to affect the structure and function of the body. To measure intent, the FDA evaluates factors like dosage guidance, marketing, and more instead of inspecting the label alone. For example, this April, the FDA issued seven warning letters against companies, like Gram Peptide,[3] selling illegally compounded peptides labelled "Research-Use-Only" and "not for human consumption," warning that these labels of stated intent provided no protection if the actual intent was clearly to affect the human body.[3, 15]

3. Ethical Concerns of the "Grey Market"

3.1 Peptide Safety Concerns and Lack of Testing and Data

Unlike peptides in category one, peptides sold on the "grey market" don't have to pass through the same rigorous testing procedures,[12] resulting in inadequate testing and data to fully prove human safety. Even if the limited testing presents health concerns, the products can remain on the "grey market" regardless, like in the case of BPC-157, which was found to potentially cause neurodegeneration through angiogenesis signaling.[16] BPC-157 had been promoted for its anti-cancer effects, among other claimed benefits, but its anti-tumor were only found in a single, nonreplicable conference abstract,[17] and there were only two human studies, with the intravenous study only testing two subjects,[18] reflecting the lack of adequate data common across these unregulated peptides. This gap is especially pronounced when searching for data on the long term effects of such peptides; no such data exists, for example, for BPC-157.

3.2 Lack of Accountability

Online vendors selling these peptides have little accountability towards their consumers,[3] and as many are based outside of the U.S., it is especially difficult for consumers to pursue legal action against them as they aren't subjected to as many of the FDA's strict regulations.[19] These counterfeit drug vendors often also lack proper documentation, or provide fraudulent records, making it difficult to trace the source of a drug or peptide.[20]

3.3 Inaccurate Formulations

Counterfeit peptides rarely follow the correct formulations. A 2025 FDA testing summary, for example, found that up to 40% of the compounded peptides contained incorrect dosages or undeclared ingredients,[21] and a 2026 MZ Biolabs Test-Buy Study found that 22% of online peptide samples contained less than 90% of their advertised peptide content.[22] Another study on the Belgian peptide "grey market" found these drugs are also easily subject to contamination from heavy metals, like lead, and industrial chemicals, and may be stored and transported incorrectly, all of which can affect

the peptides' intended performance and cause harmful side effects.[35] In addition, the products are rarely tested to check their composition, such that both the vendor and consumer are unaware if the peptide is the one it claims to be.[20]

Furthermore, for many drugs popular on the grey market, like the seven peptides, multiple different compounds are marketed under the same name.[4] Names like "TB-500" and "BPC-157" are common names, not nonproprietary names like International Nonproprietary Names (INNs) or United States Adopted Names (USANs);[36] thus, though "TB-500," for example, often refers to thymosin beta-4, a 43-amino-acid peptide, but it can also refer to different salts, active moieties, modified variants, and more.[37] Such ambiguity complicates FDA evaluation, as the different compounds under one name would have varying physiochemical and pharmacokinetic properties, and thus may not carry the same risk profile or be suitable for the same compounding use.[38]

4. Solutions and Legitimization's Impact on Safety

4.1 The PCAC's Proposed Solution

The members of FDA's PCAC, many of which are tied to the peptide industry as physicians and more, voted on several of these "grey market" peptides in July, choosing to recommend six of the seven peptides: BPC-157, MOTS-c, KPV, TB-500, Epitalon, Semax, and Emideltide. According to NPR, one of the PCAC members, Brigham Buhler, stated that this decision was to "slow down or shut down the gray and black market" by allowing these popular peptides to pass through a regulated pathway, minimizing the many safety concerns inherent in the unregulated grey market.[5] However, these recommendations contradict the advice of FDA scientists, who urged PCAC to vote against the peptides due to a lack of adequate data, like human clinical trials. This thus raises an important ethical concern: though moving these peptides to category 1 would allow them to pass through regulated pathways, reducing risk of inaccurate formulation and other errors, it could also increase access to the peptides' uncertain side effects, which could include cancer growth or a massive immune response.[5, 16] Additional ethical concerns arise from potential conflicts of interest within the PCAC. On June 29th, several months before the vote, the FDA appointed eight new temporary voting members to the advisory committee, with the majority of them connected to businesses that promote and prescribe peptides.[39] Across all seven votes, one per peptide, most of the newly appointed members consistently voted in favor of the Category 1 classification and fewer peptide restrictions.[40]

It is also important to consider that the effectiveness of legitimization is uncertain, as the "grey market" could continue to persist despite official compounding when the former offers cheaper prices, as evidenced by several drugs popular on the "grey market" today even after they were legalized.[23, 24]

4.2 Grey Market Persistence After Legalization

Legalization does not necessarily mean the "grey market" shrinks. One key example is the legalization of recreational cannabis; in California, for example, the drug was legalized in 2016, but in 2020, 80% of their cannabis was still bought and sold illegally.[23] Other states had a smaller percentage of illicit sales, with a range of 15-50% in Washington and about 13% in Colorado, but no state fell below ten percent.[23] These illegal products remained popular largely because of their lower prices. California, with a significantly higher ratio of illegal cannabis, also enacted significantly higher taxes on recreational cannabis than its state counterparts, encouraging consumers to seek cheaper options.[24]

In addition to circumventing taxes and regulations, "grey market" products are less costly because their supply chains and consumer markets are well-established and often more

efficient than their legal counterparts. Although the legal market may be safer, some consumers still buy illicit products simply because they have bought from them before, and thus view them as more trustworthy.[25]

However, "grey markets" persist both in legalization and suppression. The latter attempts to slow the illegal market through bans and stricter regulation, like warning letters,[15] with limited success just like the legalization route. Products like synthetic cannabinoids and synthetic cathinones ("bath salts"), despite emergency government scheduling prohibiting their sale and production, continued to thrive on illegal markets as manufacturers quickly modified their drugs to evade regulation.[26] In addition, a new act titled "Safeguarding Americans from Fraudulent and Experimental Drugs Act of 2025" ("SAFE Drugs Act") has proposed increased drug production facility checks, stricter limits on compounding, and other regulations to reduce the prevalence of the peptide and other experimental drugs "grey market."[13] However, given that many of the vendors are based overseas,[19] and sell through online spaces,[3] the effectiveness of this legislation on counterfeit peptides may not be as rigorous as hoped.

Nevertheless, suppression has worked, for example, in the FDA's warning letters, sent to counterfeit peptide vendors about the use of the "Research-Use-Only" label. They succeeded in shutting down seven companies which had incorrectly used the label on peptides for human consumption.[3, 15]

4.3 Safety Concerns After Legalization

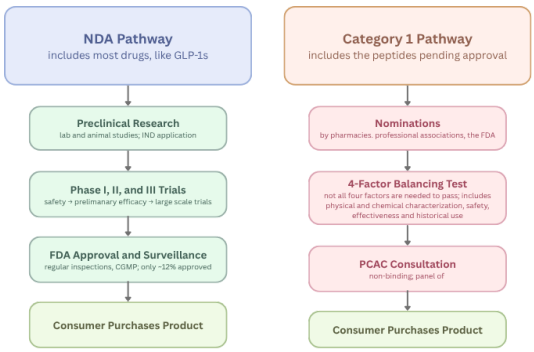


Figure 1. Category 1 designation, like in the case of the peptides, requires significantly fewer restrictions and regulations compared to the more common NDA approval that allows brand-name mass-production of the approved product.

Legalization of peptides may contribute minimal improvement to its human safety. This is especially true since peptides are regulated under the category 1, 2, and 3 system, which requires less safety testing than the traditional New Drug Application (NDA) framework that requires clinical trials and additional post-market surveillance.[28] In contrast, the FDCA category framework has more lenient requirements. It does not require additional testing nor human clinical trials, and does not require FDA inspections of those producing the peptides.[12] Furthermore, category 1 only permits compounding at local pharmacies, while the NDA includes brand-name, mass-manufactured products.

However, small-scale compounding, like unregulated "grey market" production, is more susceptible to errors, like contamination or inaccurate concentrations. A particularly infamous example is the 2012 fungal meningitis outbreak due to contaminated, compounded in methylprednisolone acetate injections from the New England Compounding Center, which caused over 750 illnesses and 60 deaths across the United States.[29, 30]

A closer example showing the risks of compounding is GLP-1s, specifically,

semaglutide and tirzepatide. Both were legitimized through the NDA framework by 2021, but officially entered shortage periods in 2022, allowing an increase in pharmaceutical compounding of the two GLP-1s. In the three years of this shortage, from 2022 to 2025, nearly 990 semaglutide adverse-event reports and over 730 associated with tirzepatide were documented, citing symptoms like severe nausea and vomiting. Of the over 1,700 reports, at least a hundred patients were hospitalized, and about seventeen passed away.[29] Nevertheless, it's important to note the FDA states that the compounding events may not have been the cause of the adverse event reports.[31]

5. Conclusion

Therapeutic peptides are an important part of medicine and a key field for innovation. However, the rigorous amount of scientific testing and evaluation needed to safely deliver these peptides to consumers cannot keep up with customer demand for these products, leading to a flourishing "grey market" further promoted by online biohacking and wellness communities. Peptides sold within the "grey market" raise several ethical concerns, from their misleading labels to inaccurate formulations and lack of accountability, all of which can pose serious health concerns to the consumer. Thus, it is imperative to find a solution to slow or stop the "grey market."

Recently, the FDA's PCAC recommended moving six of these popular "grey market" peptides, BPC-157, MOTS-c, KPV, TB-500, Epitalon, and Semax, from category 2, where production is prohibited, to category 1, potentially allowing for the production of these peptides through a regulated pathway, representing a strategy of legitimization. On the other hand, the FDA has also employed suppression against the "grey market" through warning letters, shutting down companies selling counterfeit drugs under misleading labels like "for research purposes only." Both strategies nevertheless have their issues; suppression is minimally effective against overseas companies, which make up a majority of these counterfeit peptide vendors, where regulation is harder to enforce, while legitimization risks exposing a wider audience of consumers to peptides whose long-term effects, which could include carcinogenic or immunological effects, are still uncertain.

6. References

[1] Wang, Lei et al. "Therapeutic peptides: current applications and future directions." *Signal transduction and targeted therapy* vol. 7,1 48. 14 Feb. 2022, doi:10.1038/s41392-022-00904-4

[2] Hailu, Kirubel T et al. "Unregulated Peptide Use in the Age of Biohacking: Digital Promotion, Gray-Market Access, and Emerging Public Health Risks." *Cureus* vol. 18,6 e110657. 11 Jun. 2026, doi:10.7759/cureus.110657

[3] United States, Food and Drug Administration. "Gram Peptides—721806—03/31/2026." *U.S. Food and Drug Administration*, 31 Mar. 2026, https://www.fda.gov/inspections-compliance-enforcement-and-criminal-investigations/warning-letters/gram-peptides-721806-03312026.

[4] U.S. Food and Drug Administration. "July 23–24, 2026: Meeting of the Pharmacy Compounding Advisory Committee." *FDA*, 23–24 July 2026, https://www.fda.gov/advisory-committees/advisory-committee-calendar/july-23-24-2026-meeting-pharmacy-compounding-advisory-committee-07232026. Accessed 26 Aug. 2026.

[5] Stone, Will. "FDA Advisers Vote to Ease Peptide Restrictions, Despite Agency Concerns." *NPR*, 23 July 2026, https://www.npr.org/2026/07/23/nx-s1-5903202/fda-peptides-restrictions. Accessed 26 Aug. 2026.

[6] Apostolopoulos, Vasso, et al. "A Global Review on Short Peptides: Frontiers and Perspectives." *Molecules,* vol. 26, no. 2, 2021, article 430. https://doi.org/10.3390/molecules26020430.

[7] Fetse, John, et al. "Recent Advances in the Development of Therapeutic Peptides." *Trends in Pharmacological Sciences*, vol. 44, no. 7, 2023, pp. 425–441. https://doi.org/10.1016/j.tips.2023.04.003.

[8] Ault, Alicia. "FDA Expert Panel Backs Compounding of Six Peptides." *Medscape Medical News*, 28 July 2026, https://www.medscape.com/viewarticle/fda-expert-panel-backs-compounding-six-peptides-2026a1000pgw. Accessed 26 Aug. 2026.

[9] Lee, Changhan, et al. "The Mitochondrial-Derived Peptide MOTS-c Promotes Metabolic Homeostasis and Reduces Obesity and Insulin Resistance." *Cell Metabolism*, vol. 21, no. 3, 2015, pp. 443–54. https://doi.org/10.1016/j.cmet.2015.02.009.

[10] Khavinson, V. Kh. "Peptides and Ageing." *Neuro Endocrinology Letters*, vol. 23, suppl. 3, 2002, pp. 11–144. *PubMed*, https://pubmed.ncbi.nlm.nih.gov/12374906/.

[11] Glazova, Nataliya Yu., et al. "Semax, Synthetic ACTH(4–10) Analogue, Attenuates Behavioural and Neurochemical Alterations Following Early-Life Fluvoxamine Exposure in White Rats." *Neuropeptides*, vol. 86, 2021, article 102114. https://doi.org/10.1016/j.npep.2020.102114

[12] U.S. Food and Drug Administration. *Pharmacy Compounding of Human Drug Products Under Section 503A of the Federal Food, Drug, and Cosmetic Act: Guidance*. Rev. 2, U.S. Department of Health and Human Services, June 2016, https://www.hhs.gov/guidance/sites/default/files/hhs-guidance-documents/FDA/Pharmacy-Compounding-of-Human-Drug-Products-Under-Section-503A-of-the-Federal-Food--Drug--and-Cosmetic-Act-Guidance.pdf.

[13] U.S. Food and Drug Administration. "Substances in Compounding That May Present Significant Safety Risks." *FDA*, 22 Apr. 2026, https://www.fda.gov/drugs/human-drug-compounding/certain-bulk-drug-substances-use-compounding-may-present-significant-safety-risks. Accessed 26 Aug. 2026.

[14] Turnock, Luke, and Nick Gibbs. "Click, Click, Buy: The Market for Novel Synthetic Peptide Hormones on Mainstream E-Commerce Platforms in the UK." *Performance Enhancement & Health*, vol. 11, no. 2, 2023, article 100251. https://doi.org/10.1016/j.peh.2023.100251.

[15] Policy Canary. "The 'Research Use Only' Loophole Just Closed: FDA Hits Seven Peptide Websites in a Single Day." Policy Canary, 21 Apr. 2026, https://policycanary.io/blog/fda-research-peptide-warning- letters-april-2026. Accessed 26 Aug. 2026.

[16] Józwiak, Michalina, et al. "Multifunctionality and Possible Medical Application of the BPC 157 Peptide—Literature and Patent Review." *Pharmaceuticals*, vol. 18, no. 2, 2025, article 185. https://doi.org/10.3390/ph18020185.

[17] Radeljak, S., et al. "BPC 157 Inhibits Cell Growth and VEGF Signalling via the MAPK Kinase Pathway in the Human Melanoma Cell Line." *Melanoma Research*, vol. 14, no. 4, Aug. 2004, pp. A14–A15.

[18] Lee, Edwin, and Kailynd Burgess. "Safety of Intravenous Infusion of BPC157 in Humans: A Pilot Study." *Alternative Therapies in Health and Medicine*, vol. 31, no. 5, 2025, pp. 20–24. *PubMed*, pubmed.ncbi.nlm.nih.gov/40131143/.

[19] Rathke, Richard. "Counterfeit Drug Injuries." *LegalMatch*, 6 Jan. 2022, https://www.legalmatch.com/law-library/article/counterfeit-drug-injuries.html. Accessed 26 Aug. 2026.

[20] Feeney, Aonghus J., Jeffery A. Goad, and Gerard T. Flaherty. "Global Perspective of the Risks of Falsified and Counterfeit Medicines: A

Critical Review of the Literature." *Travel Medicine and Infectious Disease*, vol. 61, 2024, article 102758. https://doi.org/10.1016/j.tmaid.2024.102758.

[21] U.S. Food and Drug Administration. "Warning Letters." *FDA*, https://www.fda.gov/inspections-compliance-enforcement-and-criminal-investigations/compliance-actions-and-activities/warning-letters. Accessed 26 Aug. 2026.

[22] Harwood, Shannon. "Gray-Market Peptide Users Face Legal Limbo as FDA Scientists Oppose BPC-157 Vote." *TechTimes*, 23 July 2026, https://www.techtimes.com/articles/321365/20260723/gray-market-peptide-users-face-legal-limbo-fda-scientists-oppose-bpc-157-vote.htm. Accessed 26 Aug. 2026.

[23] Auriol, Emmanuelle, Alice Mesnard, and Thomas Perrault. "Weeding Out the Dealers? The Economics of Cannabis Legalization." *Journal of Economic Behavior & Organization*, vol. 216, 2023, pp. 62–101. https://doi.org/10.1016/j.jebo.2023.10.006.

[24] Detrano, Joseph. "Cannabis Black Market Thrives Despite Legalization." *Center of Alcohol & Substance Use Studies*, Rutgers University, https://alcoholstudies.rutgers.edu/cannabis-black-market- thrives-despite-legalization/. Accessed 7 Sept. 2026.

[25] Goldstein, Robin, and Daniel A. Sumner. *Can Legal Weed Win? The Blunt Realities of Cannabis Economics*. University of California Press, 2022.

[26] Weaver, Michael F., et al. "The Rise of Designer Drugs: Use and Misuse of Novel Psychoactive Substances." *Addiction Science & Clinical Practice*, vol. 10, 2015, article 8. https://doi.org/10.1186/ s13722-015-0024-7.

[27] United States, Congress, House. *Safeguarding Americans from Fraudulent and Experimental Drugs Act of 2025.* 119th Congress, House Bill 6509, introduced 9 Dec. 2025, *Congress.gov*, www.congress.gov/bill/119th-congress/house-bill/6509.

[28] U.S. Food and Drug Administration. "New Drug Application (NDA)." *FDA*, https://www.fda.gov/drugs/types-applications/new-drug-application-nda. Accessed 7 Sept. 2026.

[29] Mattingly, T Joseph 2nd et al. "Adverse events administering glucagon-like peptide-1 receptor agonists: a cross-sectional study." *Health affairs scholar* vol. 4, 2023. 3 Feb. 2026, https://doi.org/10.1093/haschl/ qxag023

[30] Gudeman, Joshua, et al. "Safety of Compounded Medications." *Journal of Clinical Medicine*, 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC11796295/

[31] U.S. Food and Drug Administration. "FDA's Concerns with Unapproved GLP-1 Drugs Used for Weight Loss." *FDA*, https://www.fda.gov/drugs/drug-alerts-and-statements/fdas-concerns-unapproved-glp-1-drugs-used-weight-loss. Accessed 7 Sept. 2026.

[32] Watson, Christopher J., et al. "Pharmaceutical Compounding: A History, Regulatory Overview, and Systematic Review of Compounding Errors." *Journal of Medical Toxicology*, vol. 17, no. 2, May 2021, pp. 197–217. *Springer Nature*, https://doi.org/10.1007/s13181-020-00814-3.

[33] Zhang, Qingsong, et al. "Compounding Warning Letters to 503A Facilities Between 2017 and 2021." *Journal of the American Pharmacists Association*, vol. 63, no. 5, Sept.–Oct. 2023, pp. 1583–91. https://doi.org/10.1016/j.japh.2023.06.024.

[34] Mendias, Christopher L., and Tariq M. Awan. "Safety and Efficacy of Approved and Unapproved Peptide Therapies for Musculoskeletal Injuries and Athletic Performance." *Sports Medicine*, vol. 56, 2026, pp. 1921–35. https://doi.org/10.1007/s40279-026-02437-0.

[35] Janvier, Steven, et al. "Impurity Profiling of the Most Frequently Encountered Falsified

Polypeptide Drugs on the Belgian Market." *Talanta*, vol. 188, 1 Oct. 2018, pp. 795–807. https://doi.org/10.1016/ j.talanta.2018.06.023.

[36] Boring, Dan. "The Development and Adoption of Nonproprietary, Established, and Proprietary Names for Pharmaceuticals." *Journal of Clinical Pharmacology*, vol. 31, no. 3, 1997, pp. 312–17. https://doi.org/10.1177/009286159703100301.

[37] U.S. Food and Drug Administration. *FDA Briefing Document: TB-500-Related Bulk Drug Substances (TB-500 [Free Base] and TB-500 Acetate)*. Pharmacy Compounding Advisory Committee Meeting, 23–24 July 2026, 15 May 2026, https://www.fda.gov/media/193349/download.

[38] Elsayed, Yomnah Y., Toni Kühl, and Diana Imhof. "Regulatory Guidelines for the Analysis of Therapeutic Peptides and Proteins." *Journal of Peptide Science*, vol. 31, no. 3, Mar. 2025, article e70001. https://doi.org/10.1002/psc.70001.

[39] Lawrence, Lizzy, and Sarah Todd. "Longevity, Wellness Physicians Named to Panel Advising FDA on Peptides." *STAT*, 29 June 2026, https://www.statnews.com/2026/06/29/new-fda-peptide-advisers-compounding-committee-conflicts/.

[40] Gentilini, Arianna, et al. "Conflicts of Interest in United States Food and Drug Administration Advisory Committees: A Systematic Literature Review." *Value in Health*, vol. 29, no. 6, June 2026, pp. 1105–13. https://doi.org/10.1016/j.jval.2026.01.018.

Priced Before Proven: How Speculative Markets Interact with Scientific Discovery

Caleb Son[1*]
[1]*Mount Hebron High School, Ellicott City, MD 21042*

Abstract

Modern science relies on a lengthy and thorough chain of verification that relies on peer review and independent lab replication trials before any scientific discovery is taken as fact. In contrast, financial markets incorporate new information into prices almost instantly. Such a disparity often causes unverified scientific information to influence speculative markets. This paper argues that the resulting harm stems not from market speed alone, but from a credibility gap. The fact that a claim exists reaches markets at market speed while information about whether the claim is reliable arrives only at verification speed. This creates a disparity as most traders cannot assess credibility themselves. Using the 2023 LK-99 controversy as a case study, this paper examines how that gap produces mispricing and uncompensated losses for uninformed investors. This work concludes by proposing two targeted mitigations: a claim-specific confidence disclosure standard and a cause-triggered trading pause.

1. Introduction

In modern academia, the path between laboratory results to real-world application is laid with a multitude of verification steps. Discoveries are often first announced publicly through preprints to claim original discovery, allow free access to one's discoveries, and to gather feedback from the wider scientific community. Then, peer reviewers assess whether the methodology, evidence, and claims of the discovery are legitimate. If the work is certified to be reasonable, the work is published in a journal. Often, discoveries are vetted through independent laboratory replication of said discoveries or regulatory approval by government administrative bodies for certain domains. Finally, the discovery must garner the manufacturing and deployment resources to successfully reach market adoption.[1] The many steps in this process, which appear protracted at first glance, are argued to be essential for the safe deployment of a discovery: Open Science Collaboration showed that when 100 published psychology studies were subjected to independent replication, only 36% showed statistically significant results that matched the original findings, illustrating how crucial verification and replication steps are to achieve real scientific truth.[2] As such, scientific research revolves around organized skepticism, Merton's term for the academic institutional norm of scrutinizing a scientific claim before it is accepted as fact.[3]

Markets, in comparison, operate much faster. The Efficient Market Hypothesis, as proposed by Eugene Fama, explains that the prices of goods reflect all available information as soon as such information is made public.[4] The work makes clear that speed is not optional in markets. If a piece of news implies a good is worth more than the market price, any profit-seeking individual who notices will buy the good immediately, causing the price to quickly go up. Additionally, Hayek's classic 1945 paper, "The Use of Knowledge in Society," where he argues that speedy adjustments are the true purposes of the prices of goods in markets entirely, reveals the reasoning why a speedy market-system economy is necessary.[5] Prices are the coordination mechanisms that society uses to

produce what it needs; by quickly raising or lowering the price of a stock or good, markets are able to effectively signal throughout society where scarce capital must be allocated most.

Thus, a disparity in processing speed is clear between markets and academia. Fast market reactions, driven by the same economic incentives that keep society afloat, take signals from unconfirmed preprints, leaks, and unreplicated papers, causing fluctuations in prices from speculative and unreliable information. This can trigger losses for retail investors and misallocation of capital across the economy, hurting both average consumers financially and society as a whole, even if the misinformation from the scientific discovery was spread without intent to deceive. One of the clearest examples of this phenomenon arose in the summer of 2023, when a team of South Korean researchers at the Seoul-based Quantum Energy Research Center (Q-centre) and Korea University submitted two preprints on a newly discovered material they named LK-99.[6,7] The work claimed LK-99 to be the world's first ambient-pressure room-temperature superconductor; throughout the summer, stock prices for energy technology corporations soared as social media brought LK-99 into mainstream public knowledge and news.[8] Weeks later, failures of independent labs to replicate results not only publicly debunked LK-99 but also triggered significant economic ripples as markets adjusted to the newfound information.[9,10]

LK-99 is a useful case study for this problem because its entire lifecycle is well documented and it achieved real significant impact. The claim went from an obscure preprint to recognition across multiple news platforms to a scientific consensus of failure in less than a month, with each stage leaving a clear public trail across social media, journal papers, and the news. Additionally, the recency of the case combined with its virality confirms the modern relevance and wider impact of the case, distinguishing it from both older cases and smaller-scale cases that have weaker documentation of economic impact.

The central problem is not that markets are fast. The main cause lies in a credibility gap. News that a claim exists travels at market speed, while evidence of whether the claim is credible travels at verification speed. In the interval between the two, most market participants lack the expertise to judge the claim themselves. Thus, prices come to reflect attention and excitement rather than the probability that the claim is true. This mispricing imposes losses on investors and distorts capital allocation. This paper uses the LK-99 episode to trace this mechanism, evaluate its ethical implications, and propose two mitigations: a disclosure standard that supplies the missing credibility signal and a trading pause that limits herding when that signal is absent.

2. Case Study: LK-99

Since the early 1900s, the field of materials science has been searching for an ambient room-temperature superconductive material, technically defined as a material capable of conducting electricity with exactly zero resistance. Such a material, if found, would have the potential to negate nearly all energy loss in power grids and make advanced electronics and computing hardware much more efficient.[11] Additionally, the ability to run electric current through superconductors without energy loss in heat allows for extremely strong and cheap electromagnets via the Meissner Effect, which would revolutionize transportation, nuclear reactors, medical imaging machinery, and other high-power magnet technologies.[12] Overall, such a material would likely be hailed as one of the greatest physics discoveries of the century, as any technology

relying on electrified systems or magnetic power would instantly be revolutionized.

Typically, superconductive materials only reach zero resistance once cooled below a specific critical temperature known as Tc. Lower temperatures cause atom vibrations and lattice disruptions to be minimized, allowing electron pairs to glide smoothly through the material without interference, and thus, without any energy loss. However, these critical temperatures can often reach depths of nearly -133°C, requiring bulky cooling equipment worth millions of dollars.[11] While past research has made great strides in raising the Tc of superconductive materials, even the most advanced materials require temperatures far below room temperature to show superconductivity. This currently keeps superconductive materials allocated to only a few specialized applications in MRI imaging and particle accelerator research laboratories.

On July 22, 2023, two preprints from a South Korean energy research team were posted on arXiv, an open-access academic archive with no peer-review requirement for posts, claiming to have discovered a superconductor with a critical temperature well above room temperature. The material, dubbed LK-99 by the researchers, was a copper-doped lead apatite that demonstrated a sharp drop in electrical resistivity and partial magnetic levitation over a permanent magnet, presented as the Meissner effect. The researchers hypothesized in the preprints that localized regions of superconducting behavior were created from copper atoms substituting for lead atoms in the material's lattice.[6] Within days, social media caused LK-99 to go viral, where it quickly gained attention amongst both academic circles and the general public.

Within a few days, both domestic and international markets had reacted to the hype around LK-99. Despite no peer verification or lab replication of the preprints' results, speculative trading swept through markets almost immediately. Domestically, small-cap stocks from Duksung Co. and Sunam Co., two prominent manufacturing companies based in South Korea, saw incredible leaps, with Sunam climbing 260% over 6 trading days and Duksung rising 170% total.[13] This effect could also be noted internationally. In China, Jiangsu Fasten Co. and Henan Zhongfu Industry Co. also saw prominent spikes in shares during LK-99's reveal and consequent slumps once independent debunkings of the material began to emerge.[14] Additionally, American Superconductor Corporation, an energy technologies company, also saw a record 60% surge in stocks during the earlier stages of the LK-99 incident, a jump that was more than 43 times its stock's average daily trading volume.[15] Interestingly, the corporation had no business ties or relation to LK-99 and stood to lose a large amount of potential revenue if a genuine room-temperature superconductor had actually reached the market through a competitor. It appears that the jump in stock was triggered by investors primarily reacting to the company's name, including the term 'superconductor', which illustrates how little scientific literacy is required for a claim to move real money and demonstrates the fragility of speculative market interactions with the scientific community.

As LK-99 became more widespread, members of the scientific community began reviewing and attempting to replicate the preprints' results. Initially, many scientists expressed skepticism at the idea of LK-99 being the world's first ambient-pressure room-temperature superconductor, with critics noting inconsistent evidence and reasoning in the South Korean team's explanation for LK-99's superconductivity.[6] On August 8, 2023, about two and a half weeks after the initial reporting of LK-99, researchers at the University

of Maryland announced that LK-99 showed no exceptional superconductive properties.[16] Various other labs utilized theoretical models and wet-lab replication efforts to also indicate the lack of exceptional superconductivity in LK-99.[17,18,19]

While LK-99 remains a failed attempt at an age-old scientific question, the interaction between the supposed discovery and speculative markets revealed that scientific research can have a real and noticeable impact on both society and markets. Additionally, the timing of the public and market reactions occurring well before replication trials or the peer-review process clearly shows the structural disparity between markets and the scientific process.

It is imperative to emphasize that LK-99 is not an outlier case of such a disparity being on display. In 1989, a pair of University of Utah chemists announced at a press conference, rather than through a peer-reviewed paper, to have achieved room-temperature nuclear fusion.[20] This led to the Utah State Legislature convening a special session and appropriating nearly $5 million in public funds, towards establishing the National Cold Fusion Institute; after the discovery was found to be false, the Institute was promptly dissolved. Similarly, palladium, a material used in the chemists' experiment, saw a sudden surge in its prices across the nation.[21] A similar case also occurred in the early 2020s, when unverified claims from the Journal of Neuroscience in a data manipulation scandal involving biotech firm Cassava Sciences caused stocks in the company to quickly rise nearly 76%.[22,23] Across both of these cases, a pattern emerges: an unverified scientific claim reaches the ear of both the public and the market, causing surges or drops in stocks and fund allocations before even the first steps in the scientific verification process is undergone. Overall, the prevalence of this disparity between markets and the scientific research process raises two concerns: its ethical implications, and how it might be resolved.

3. Ethical Analysis and Discussion

In isolation, both science and markets work ethically and effectively; academia utilizes extensive review to protect against premature belief, while markets protect against inefficient capital allocation. Once each system begins interacting with the other, however, an issue of information asymmetry, where one party to a transaction or decision holds more or better information than another, arises; markets simply lack the scientific expertise necessary to understand the likelihood of early-stage discoveries yielding results. Announcing a result publicly before any academic verification through successful replication trials or peer review results in a negative externality, an economic term for a cost imposed on an uninvolved third party who did not consent to the decision that caused it. In this case, investors and the public bear an unexpected cost from decisions they have no say in. This asymmetry also creates a moral hazard problem: scientists or insiders with the expertise to see through false scientific claims could unfairly manipulate markets for profit before the real truth becomes public. Although LK-99 itself showed no evidence of this occurring, the structural opening for it to potentially occur reveals the ethical danger of the disparity's existence.

Two other mechanisms also compound to this. Firstly, even in a world with perfect transparency, the pure amount of scientific complexity new discoveries hold means most market participants cannot evaluate a claim like LK-99 on its merits. Additionally, subtle unethical behavior can also arise. Researchers, firms, or promoters may strategically avoid disconfirming information or present real but misrepresentative information as truth. These can all arise through methods like showing

selective results, favorable framing, or exaggerated implications. Thus, these two reasons emphasize why a disclosure label alone cannot be the whole answer; a solution would require intervention from expert involvement, with a substantial background in field that the discovery was made in.

From a consequentialist lens, the risk the current friction between the two systems poses outweighs the benefits. LK-99's public spread caused accelerated scrutiny and faster independent replication attempts as more economic incentive was given for the truth behind LK-99 to be revealed, as more investment money poured into conductive materials and electronics markets. However, the accelerated scientific focus only truly pays off if the claim in question is true, something completely unknowable at the moment. Every time a widespread claim is proven false, late-arriving retail investors absorb losses while the larger economy bears the cost of misallocated capital diverted away from productive use. This issue can occur even without malicious intent from the scientist's perspective. As seen in the case of LK-99, the South Korean research team behind the original discovery had no apparent intention of misleading the market for personal gain, revealing this as an issue stemming purely from structural friction.

It is important to emphasize that existing safeguards did not prevent the LK-99 episode. Preprint servers such as arXiv state that postings are not peer reviewed, yet this label is generic: it appears identically on every submission and conveys nothing about how credible a particular claim is. A reader cannot distinguish a speculative claim from a routine one. The Korea Exchange did issue a warning about the superconductor rally, but it arrived after the rally was already well underway, emphasizing the need for a faster institutional mechanism. Meanwhile, standard circuit breakers are triggered by the size of a price movement rather than its cause while also lasting only minutes. This makes them far too short for scientific verification and too blunt to distinguish a science-driven surge from any other volatility.

Because both academic and market systems still work separately, solutions would need to revamp how the two systems interact with each other rather than internally shifting the design of either mechanism. On the academia side, a large-scale confidence disclosure standard that any unverified claim made public must bear could help clarify the reliability of a result or discovery made public. The standard, drawing inspiration from existing journal embargo systems, could be removed after a sufficient level of review and replication has occurred to back a claim.[24] On the market side, a circuit breaker extension for research-related volatility could help markets step back from scientific claims until verification is confirmed. Realistically speaking, such a mechanism could be overseen by an exchange staff subcommittee composed of scientific experts and statisticians that halt market shifts for a defined period of time when they determine that an abnormal price change is triggered by unverified scientific claims.

These solutions are not perfect; labels can be ignored by readers, and circuit breaker extensions cannot last for the months required to properly and thoroughly verify a scientific claim. However, both allow the two worlds to interact with greater caution and awareness, lessening the friction between markets and academia and the ethical burden of the issue of miscommunication of crucial scientific information.

Additionally, it is important to acknowledge that a scientific claim can disappoint markets in more than one way. It can be false, as LK-99 was. Yet, it can also be true

while being slower or narrower to commercialize than investors initially assume due to scaling difficulties, cost, regulatory hurdles, or patent and priority disputes. This can carry unintended consequences that reduce its value. Materials such as graphene and carbon nanotubes are real and well characterized, yet their commercial timelines have repeatedly lagged behind early enthusiasm. The proposals in this paper primarily address the first failure type, in which the underlying claim lacks verification. Greater market and academic infrastructure would be required to address this second failure type.

4. Conclusion

The excitement and misinformation surrounding LK-99 has all but died down, as the various failed verification efforts emerged, one by one, in both the news and academic journals. The search for a convenient superconductive material continues. Yet, the lessons gleaned from studying LK-99 reveal a key structural flaw in the interaction between markets and academia. A credibility gap exists, as claims reach markets at market speed while evidence of their reliability arrives at verification speed, leaving non-expert traders to price a claim they cannot evaluate. As markets grow faster and more automated, this gap will not be closed on its own. Additionally, many fields are racing towards their own ‘room-temperature superconductor breakthrough’, whether that be in quantum computing, gene editing, or artificial intelligence, subsequently causing every field to be vulnerable to the unresolved gap identified in this paper. Narrowing that gap, even partially, will require both worlds to work together to build the infrastructure necessary to prevent the next viral scientific claim from growing out of control.

5. References

[1]University of California Museum of Paleontology. "Science Flowchart - Understanding Science." *Understanding Science*, University of California, Berkley, 25 Apr. 2022, https://undsci.berkeley.edu/science-flowchart/.

[2]Open Science Collaboration. "Estimating the Reproducibility of Psychological Science." *Science*, vol. 349, no. 6251, Aug. 2015, https://doi.org/10.1126/science.aac4716.

[3]Merton, Robert King. *The Sociology of Science Theoretical and Empirical Investigations*. University of Chicago Press, 1993.

[4]Fama, Eugene. "Efficient Capital Markets: A Review of Theory and Empirical Work." *The Journal of Finance*, vol. 25, no. 2, May 1970, pp. 383–417, https://doi.org/https://doi.org/10.2307/2325486.

[5]Hayek, Friedrich. "The Use of Knowledge in Society." *The American Economic Review*, vol. 35, no. 4, 1945, pp. 519–30, http://www.jstor.org/stable/1809376.

[6]Lee, Sukbae, et al. "The First Room-Temperature Ambient-Pressure Superconductor." *arXiv.Org*, 22 July 2023, https://arxiv.org/abs/2307.12008.

[7]Lee, Sukbae, Jihoon Kim, Hyun-Tak Kim, et al. "Superconductor $Pb_{10-x}Cu_x(PO_4)_6O$ Showing Levitation at Room Temperature and Atmospheric Pressure and Mechanism." *arXiv.Org*, 22 July 2023, https://arxiv.org/abs/2307.12037.

[8]Zilber, Ariel, and Shannon Thaler. "LK-99 Superconductor Research Breakthrough Could Mark 'New Era for Humankind.'" *New York Post*, 2 Aug. 2023, https://nypost.com/2023/08/02/lk-99-superconductor-breakthrough-could-mark-new-era/.

[9]Garisto, Dan. "LK-99 Isn't a Superconductor — How Science Sleuths Solved the Mystery." *Nature*, vol. 620, nos. 705–706, Aug. 2023, https://doi.org/10.1038/d41586-023-02585-7.

[10]Bloomberg. "Energy Connects." *Energy Connects*, 8 Aug. 2023, https://www.energyconnects.com/news/utilities/2023/august/doubts-about-lk-99-superconductor-claim-sink-korean-stocks.

[11]Ball, Philip. "There's No Room-Temperature Superconductor yet, but the Quest Continues." *The Guardian*, 2 Sept. 2023, https://www.theguardian.com/science/2023/sep/02/room-temperature-superconductor-south-korea-lk-99-nuclear-fusion-maglev.

[12]Hirsch, J. E. "The Meissner Effect in Superconductors: Emergence Versus Reductionism." *Journal of Physics: Condensed Matter*, vol. 38, no. 7, Feb. 2026, p. 073001, https://doi.org/10.1088/1361-648x/ae3cf2.

[13]Lee, Youkyung. "Superconductor Mania Triggers Korea Exchange Warning." *Yahoo Finance*, Yahoo! Finance, 4 Aug. 2023, https://finance.yahoo.com/news/superconductor-stock-mania-triggers-warning-024638977.html.[11]

[14]Yu, Jeanny, and Youkyung Lee. "Superconductor Stocks' Wild Swings Are Getting Risky." *Yahoo Finance*, Yahoo! Finance, 14 Aug. 2023, https://finance.yahoo.com/news/superconductor-stocks-wild-swings-getting-025500885.html.

[15]Bloomberg News. "Superconductor LK-99 Breakthrough Buzz Spurs China, Korea Tech Rally." *Bloomberg.Com*, Bloomberg, 2 Aug. 2023, https://www.bloomberg.com/news/articles/2023-08-02/superconductor-breakthrough-buzz-spurs-china-korea-tech-rally.

[16]Paleja, Ameya. "LK-99: Replicated Material Does Not Show Superconductivity." *Interesting Engineering*, 9 Aug. 2023, https://interestingengineering.com/science/lk-99-replicated-material-no-superconductivity.

[17]Griffin, Sinéad M. "Origin of Correlated Isolated Flat Bands in Copper-Substituted Lead Phosphate Apatite." *arXiv.Org*, 3 Aug. 2023, https://arxiv.org/abs/2307.16892.

[18]Kumar, Kapil, et al. "Absence of Superconductivity in LK-99 at Ambient Conditions." *ACS Omega*, vol. 8, no. 44, Oct. 2023, pp. 41737–43, https://doi.org/10.1021/acsomega.3c06096.
[19]Habamahoro, T, et al. "Replication and Study of Anomalies in LK-99—the Alleged Ambient-Pressure, Room-Temperature Superconductor." *Superconductor Science and Technology*, vol. 37, no. 4, Feb. 2024, p. 045004, https://doi.org/10.1088/1361-6668/ad2b78.
[20]Reid, T. R. "Utah Grants $5 Million to Fusion Experiments." *The Washington Post*, 8 Apr. 1989, https://www.washingtonpost.com/archive/politics/1989/04/08/utah-grants-5-million-to-fusion-experiments/7bf760ce-e05b-48fb-bd63-ced32716cfa5/.
[21]Associated Press. "FINANCIAL MARKETS : Commodities : Fusion Tests Lift Palladium Prices to Highest Point Since '83." Los Angeles Times, 13 Apr. 1989, https://www.latimes.com/archives/la-xpm-1989-04-13-fi-1865-story.html. Accessed 10 Sept. 2026.
[22]Lenihan, Rob. "Cassava Stock Soars; Review Finds 'No Evidence of Manipulation.'" *TheStreet*, 4 Nov. 2021, https://www.thestreet.com/investing/cassava-sciences-stock-soars-as-review-finds-no-evidence-of-manipulation.
[23]Globe Newswire. "Review by Journal of Neuroscience Shows No Evidence of Data Manipulation in Technical Paper Foundational to Cassava Sciences' Lead Drug Candidate; It States, 'No Evidence of Data Manipulation Was Found for Western Blot Data..'" *BioSpace*, 4 Nov. 2021, https://www.biospace.com/review-by-journal-of-neuroscience-shows-no-evidence-of-data-manipulation-in-technical-paper-foundational-to-cassava-sciences-lead-drug-candidateit-states-no-evidence-of-data-manipulation-was-found-for-western-blot-data.
[24]SMC Spain. "What Are Embargoes? How Do Journals and Peer Review Work." *SMC España*, 23 Mar. 2022, https://sciencemediacentre.es/en/what-are-embargoes-how-do-journals-and-peer-review-work.

Journal of Ethics in Scientific and Technological Innovations

A Mammoth Undertaking: Ethical Limits on the Commercialization of De-Extinction

Lydia Choi[1*], Bella Lin[2*], Lelin Tsai[3*], Andrew Wang[4*]
[1]*Korea International School, Gyeonggi-do, South Korea,* [2]*Island Trees High School, Levittown, New York,*
[3]*Greenwich High School, Greenwich, Connecticut,* [4]*Riverdale Country School, Bronx, New York*

Abstract

In April 2025, Colossal Biosciences announced the birth of three gene-edited gray wolf pups marketed as the first successfully "de-extinct" dire wolves, bringing a long-running scientific and ethical debate over whether extinction can, or should be reversed, into the greater public. This paper examines de-extinction, the process of creating an organism that is, or closely resembles, a member of an extinct species, through back-breeding, cross-species cloning, and CRISPR-based genetic engineering, using the dire wolf, woolly mammoth, and Pyrenean ibex as case studies. Drawing on bioethics, conservation biology, and legal scholarship, this paper argues that de-extinction raises five principal ethical concerns: harm to surrogated and “revived” animals, the ecological displacement of organisms that may be returned to ecosystems no longer capable of supporting them, a moral hazard in which the existence of technology capable of de-extinction reduces the perceived urgency of preventing it, ambiguities surrounding species identity and commercial spectacle, and the field’s uncomfortable proximity to human-ancestor de-extinction. A comparative case-study analysis demonstrates that these various concerns do not weigh equally across projects: the dire wolf case reveals commercial incentives outpacing both scientific and ecological justification, while the mammoth project leans on an untested ecological restoration claim. Furthermore, this paper finds that no binding, de-extinction-specific regulatory framework currently exists; oversight is instead assembled and applied unevenly from advisory guidelines, ones never designed for engineered organisms, and patent law doctrines created for an unrelated purpose. This paper concludes that de-extinction is too varied to be judged as a single technology. Because its welfare, ecological, and commercial stakes differ sharply from project to project, it should be governed by binding, case-by-case, multidisciplinary external review, with a firm prohibition on human-adjacent research and a funding priority for living endangered species over extinct ones. This framework is a more defensible path than either commercial self-regulation or a blanket policy.

1. Introduction

In April 2025, Colossal Biosciences announced that three gray wolf pups, each carrying 20 edits across 14 genes, were the world's first "de-extinct" dire wolves.[1, 2] The pups were born between October 2024 and January 2025, but the world learned of them through a press release rather than a peer-reviewed study, and no binding external review preceded or followed the announcement. Within weeks, the "revival" was cited by U.S. policymakers in arguments for weakening Endangered Species Act protections for gray wolves.[5, 6] The episode illustrates the gap this paper takes as its starting point: de-extinction, the process of reviving an extinct species or creating one that is greatly similar to it,[3] has moved from academic speculation to a well-funded commercial sector faster than any institution has developed the authority to oversee it.[5, 6, 9]

Most public debate asks whether de-extinction is ethical in general. This paper argues that the question is poorly posed. The term covers technologies as different as back-breeding zebras, cloning a wild goat subspecies, and multiplex CRISPR editing of a wolf genome,[4, 14] and the ethical stakes differ just as widely across them. The more productive question is how a technology this varied should be governed, and by whom.

To answer it, this paper first describes the main de-extinction methods (Section 2). It then identifies five principal ethical concerns: species identity and commercial spectacle, animal welfare, ecological displacement, moral hazard, and the field's proximity to human-ancestor de-extinction, and considers the potential medical benefits sometimes offered in de-extinction's defense (Section 3). Next, it applies a four-part rubric (restorative justice, animal welfare cost, ecological readiness, and resource allocation) to three projects at very different stages: the Pyrenean ibex, the dire wolf, and the woolly mammoth (Section 4). The three cases produce three different verdicts, and that divergence is this paper's central finding. Sections 5 and 6 weigh binding oversight, case-by-case evaluation, and self-regulation, and conclude that the evidence supports neither blanket endorsement nor blanket prohibition. It supports binding, case-by-case, multidisciplinary external review, a firm prohibition on human-adjacent research, and a funding priority for currently endangered species, a priority grounded in the one clearly beneficial result of the dire wolf project, its incidental cloning of endangered red wolves.[1, 10]

2. How De-Extinction Works

De-extinction has been achieved through several strategies, each resulting in a different relationship between the revived organism and the species it is replacing.

2.1 Back-Breeding

The oldest technique is known as back-breeding, in which living relatives are selectively bred over many generations to reintroduce lost traits.[4, 14] This is not a literal reversal of evolution, but rather the recovery of a lookalike from genetic variation that remains present in the gene pool. The earliest example of back-breeding can be found in 1920s and 1930s Germany, which produced the Heck cattle and tarpan breeding programs, aiming to recreate the aurochs and wild horse from domestic cattle and horse breeds.[15] A later, more well-documented endeavor is South Africa's Quagga Project, launched in 1987, involving the selective breeding of plains zebras for quagga-like features, including reduced striping.[16, 17] Results were visible by the fifth generation; in 2005, offspring were noticeably displaying recognizable quagga traits.[16, 17] That being said, back-breeding is the slowest and least technologically demanding of the three methods. It, however, does produce the least ambiguous result: it is clear that the "de-extinction" seen through this process is a close approximation, not a genetic restoration.[14, 18]

2.2 Cross-Species Cloning

Cross-species cloning is a laboratory technique in which the nucleus from a donor cell of one species is transferred into an enucleated oocyte (an egg cell with its own nucleus removed) from a different species.[4, 19] It uses preserved tissue or ancient DNA, often extracted from museum specimens, permafrost remains, or biobanked cell lines, to generate an embryo carrying the extinct species' genome.[3, 10] Due to frequent degradation of ancient genetic material, this genome may need to be "repaired" through genetic engineering before cloning is viable.[20, 21]

Once a live individual is produced, subsequent generations can be back-bred to more closely resemble the extinct ancestor over time. The mammoth de-extinction project illustrates this combined approach. It begins with an elephant egg carried by an elephant surrogate; then relies on generations of selective breeding to gradually suppress elephant characteristics while amplifying mammoth ones.[8, 10] However, the clearest historical example of cross-species cloning is the 2003 Pyrenean ibex, where scientists used frozen skin cells preserved from the last living female to generate cloned embryos, which were implanted into domestic goat eggs and carried to term by goat surrogates.[22] It remains the only case in which a cloned extinct animal has been born alive.[14, 22]

2.3 Genetic Engineering

The most modern and technical strategy relies on the direct construction of a genome by synthesizing or editing DNA using technologies like CRISPR-Cas9.[5, 10] With these technologies, scientists can introduce the genetic variants of an extinct species into a living relative's genome, without having to rely on preserved cells to clone from directly.[4] Colossal Biosciences' dire wolf project followed a five-step process to execute this method, which is now the industry standard.[1, 2] First, well-preserved ancient DNA samples must be recovered and sequenced. Next, the target genes of the ancient species must be identified through comparative genomics, functional genomics, and prior literature, and they can be validated through organoid or in-vitro models.[10] Then, using living cells as a base, CRISPR is used to insert designed multiplex edits, eventually creating a stable cell line which can be validated for both intended edits and any unintended side effects.[1, 2] Then, an embryo is transferred via a surrogate of a closely related living species, and the new species is born. Lastly, the new species is bred in captivity, then eventually reintroduced to the wild. This step requires intensive monitoring and preparation, best accomplished through coordination with governments and conservation agencies.[18, 23] This step has yet to be accomplished by any de-extinction project.[6, 10]

For the dire wolf, this process involved the editing of 20 gene loci across 14 genes in gray wolf cells to introduce the traits of dire wolves, followed by gestation in a domestic dog surrogate.[1, 2] 14 edited genes out of ~20,000 total make this new animal over 99% gray wolf, despite the dire-wolf-like traits.[24, 25] This is an essential distinction to be made with genetic-engineering-based de-extinction, with implications in how we should treat these animals, as well as how these animals would function, should they be reintroduced into the wild.[14, 18]

2.4 Advanced Reproductive Technologies and Biobanking

Several enabling technologies sit underneath all three principal strategies. With induced pluripotent stem cells (iPSCs), researchers can bypass the need for live eggs or embryos from rare or extinct-adjacent species.[10] In vitro gametogenesis, commonly referred to as IVG, then pushes this further by growing sperm and egg cells directly from stem cells in a lab environment, particularly useful in cases akin to the near-extinct northern white rhino, of which the only remaining members of its species are two non-reproductive females.[10, 26] Another technique, which has remained unreliable, is interspecies gestation, in which a female carries an embryo of a different species in her womb. This has produced successful trials including the gaur, mouflon, wildcat, and grey wolf.[27, 28, 29, 30] However, fatal trials remain, with one high-profile rhino surrogacy attempt failing due to an infection pre-delivery.[26] Fully artificial wombs have been researched and explored since

the 1950s, but have never achieved full-term gestation in any species.[31]

None of these technologies would be possible without biobanking. The San Diego Frozen Zoo, one of the foundational biobanks, holds over 10,000 tissue and genetic samples across roughly 1,000 taxa, and is now part of a larger international repository network.[32, 33] Notably, biobanked material does not merely consist of extinct animals; rather, much of its conservation value lies in the preservation of genetic diversity from species that are still alive but declining, which is part of why several of the sources in this paper's institutional analysis argue biobanking has value independent of whether de-extinction itself ever succeeds.[6, 10]

3. Ethical Analysis

This section examines the five principal ethical concerns raised by de-extinction. Species identity and commercial spectacle are treated in two subsections (3.1 and 3.2), followed by animal welfare (3.3), ecological displacement (3.4), moral hazard (3.5), and the human limit (3.6). Section 3.7 then considers the potential medical benefits sometimes offered as a counterweight to these concerns.

3.1 Species Identity

One of the most foundational ethical problems with de-extinction concerns species identity: none of the current methods produce an organism that is entirely equivalent to the extinct species, raising the question of what is actually being created and the obligations, if any, humans owe to it.[34, 35] Back-breeding produces animals selectively bred to resemble an extinct ancestor; cross-species cloning produces hybrids; and modern genetic engineering edits only a small fraction of a living relative's genome to introduce the traits of the extinct species.[4, 14] Colossal Biosciences' dire wolf efforts illustrate just how limited this resemblance actually is. The process itself edited only 20 gene loci across 14 genes in gray wolf cells, out of roughly 20,000 total genes, meaning the resulting animal remains over 99% gray wolf despite displaying dire-wolf-like traits.[2, 25] This distinction is essential to the ethical status of the organism, since a creature that is genetically almost indistinguishable from its living relative cannot straightforwardly inherit the ecological role, behavioral instincts, or evolutionary history of the species it is meant to represent.[6, 24, 36]

Philosopher Alastair Gunn, known for his work in environmental ethics and the moral status of animals, argued that once a species becomes a "null class" following the death of its last member, that class can never regain members by definition.[34, 37] On this view, de-extinction does not revive the original species at all, but instead produces an entirely new, human-made organism artificially engineered to occupy the identity of one that no longer exists.[3,35]

3.2 Commercialization

Beyond questions of species identity, de-extinction raises a separate ethical concern rooted in commercialization: whether it is acceptable to engineer novel traits into an existing, non-endangered species, like the gray wolf, market the result as something it is not, like the dire wolf, and profit from the spectacle of a "resurrected" animal that functions largely as commercial branding rather than genuine ecological restoration.[3, 6] This commercial aspect complicates the ethics further because decisions about which traits to amplify, such as the dire wolf pups' notably oversized structures, may be driven less by necessity and more by public attention.[2, 38]

In addition to the emphasis on certain characteristics of extinct species, Colossal Biosciences' selection of target species

reinforces the concern of commercialization as the company has prioritized large, visually striking animals like the mammoth and giant moa over smaller, arguably more ecologically significant candidates such as the dodo or thylacine, suggesting that commercial appeal, rather than restoration value, may be driving which species are chosen in the first place.[2, 6, 38]

This commercial dimension intersects with an underexplored regulatory pathway: patent law. Because de-extinct organisms are technically patentable inventions, their commercialization is subject to the same balancing test used for other genetically engineered lab animals, such as the OncoMouse (the first animal ever granted a U.S. patent, genetically modified to carry an activated cancer-causing gene), which weighs animal suffering and environmental risk against the invention's usefulness to humanity.[8, 39] As a result, patent law may end up functioning as a practical checkpoint on de-extinction commercialization, one that operates entirely separately from wildlife or biosafety law.[9, 40] Scholars analyzing this pathway conclude that it is likely to accommodate de-extinction with relatively few barriers, since patent offices have historically weighed public engagement and conservation-awareness value alongside direct scientific benefit.[8, 41] This raises a further question worth considering: whether patent law, a legal framework built for evaluating inventions, is an appropriate way to adjudicate the ethical questions raised by de-extinction at all.[8, 42]

3.3 Animal Welfare

The most pressing ethical cost of de-extinction falls on the individual animals, not the species or environment.[7, 43] The clearest example is the 2003 revival of the Pyrenean ibex. The clone died within minutes of birth due to a malformed lung, a defect that had been documented in the early cross-species cloning of sheep.[19, 22] Over two decades of technical progress have not eliminated this procedural risk.[8, 10] The mammoth project raises further concerns: because Asian elephants are themselves endangered, using them as egg donors and surrogates risks the lives of living endangered creatures in the pursuit of reviving another.[8, 44] Moreover, reviving an old species of elephant comes with the issue of introducing a founder population of elephants without the herd structure they rely on to survive.[7, 45]

3.4 Ecological Displacement

Even a genetically successful de-extinction faces a greater hurdle: the ecosystem the organism once lived in may cease to exist.[46, 47] Habitats, of course, change over centuries or millennia, often for the same reasons (climate change, human interference, land use) that caused the organism's extinction originally.[36, 48] Therefore, successfully revived organisms are often unable to return to a familiar, native environment, and are instead confined in an artificial setting, including zoos, labs, or fenced preserves, which critics have argued in and of itself is a defining characteristic of bio-objectification rather than genuine restoration.[3] That is, many cases end as spectacles.

Science writer Elizabeth Preston's concept of "functional extinction" further develops this point: species loss often also leads to the loss of an ecological arrangement of relationships between organisms, rather than just the loss of an individual species.[21, 46] As such, from this perspective, even a technically successful revival can fail to restore what was actually lost: the web of predator-prey relationships and habitat engineering that give the species its ecological role may be gone regardless of whether the organism itself return.[47, 49] This is precisely the critique leveled

at Colossal's dire wolf and mammoth projects: reintroducing an apex predator or an arctic ecosystem engineer presumes we understand the precise ecological role they would play today, when in fact neither the supporting food web nor the animal's modern niche is well understood.[6, 36] Unlike historical cases of invasive species, where ecological risk wasn't recognized nor accounted for until the damage was already caused, de-extinction offers regulators the unique opportunity to assess these risks before release - an opportunity critics argue is being wasted, since species selection for de-extinction seems to be more driven by commercial appeal than planned ecological utility.[6, 38, 18]

3.5 Humanity's Moral Hazard

Humanity's participation in the revival of extinct species must also be heavily considered. If de-extinction becomes readily achievable, the public and its policymakers may begin to lose urgency for the prevention of extinction.[11, 12] This is not just a hypothetical problem. Within weeks of Colossal's announcement of their revival of the dire wolf, it was cited by lawmakers to weaken the Endangered Species Act protections for gray wolves, demonstrating how the risk of lost urgency translates into policy consequences almost immediately.[5, 6]

There is also the moral and philosophical concern of whether humanity has the right to take control of nature through de-extinction. The revival of extinct species was quickly used for symbolic and commercial purposes rather than true ecological restoration, placing the role of these animals in the world entirely in the hands of humanity. Critics worry that humanity would gain essentially full control over which species exist, an objection often summarized as "playing god".[50, 51] However, this argument has been debated: under scrutiny, many of our conservation tools today—captive breeding programs, ecological monitoring, and GPS tagging—would collapse under the same moral objection applied to de-extinction, and yet these tools are widely considered morally acceptable.[34, 50] Still, there may be a distinction between the tools. Should we treat monitoring and preserving a species differently from engineering one into existence?

3.6 The Human Limit: Neanderthal De-Extinction as a Stress Test

Nearly every expert and source in this field, regardless of stance on de-extinction as an idea, treats one hypothetical as fundamentally unacceptable: applying the same techniques to human relatives.[5, 52] Since Svante Pääbo's team sequenced a significant portion of the Neanderthal genome in 2010, the foundation for a Neanderthal "de-extinction" has technically existed.[13] Bioethicists who would otherwise disagree vehemently on the acceptability of animal de-extinction agree here: one prominent analysis argues that the recreation of a Neanderthal is unacceptable, using the comparison of human cloning as a scientifically uncertain experiment on an unconsenting surrogate mother.[5, 52] Some scholars go further, arguing that the fact that current de-extinction technologies could be plausibly extended to Neanderthals means that the field needs strict international regulation before a private company attempts it, rather than act in a reactionary manner as regulators have historically done with other biotechnologies.[5, 6]

This near-consensus is worth pausing on, as it provides some clarification on what is actually at stake in the rest of this paper's argument. If everyone agrees that genetic authorship should stop at the boundary of our own species, that agreement suggests some principle is doing the work - perhaps consent, perhaps the moral weight of suffering, perhaps the line between honoring the dead and

manufacturing the living. Whatever it is, this paper argues it does not apply only to Neanderthals or Denisovans. Those same concerns about consent, welfare, and the limits of treating a living being as an engineered product apply, in a smaller but very real degree, to the dire wolf pups and the mammoth-elephant hybrids already being born. The Neanderthal hypothesis should not be treated as a separate issue, for the underlying concerns should be applied universally to other "revived" organisms.[5, 43]

3.7 Potential Medical Benefits

Proponents sometimes point to medical benefits as an additional justification for de-extinction. The evidence suggests that these benefits come from the underlying tools and biology rather than from reviving extinct species themselves, which affects how much weight they should carry in evaluating individual projects.

Several overlaps exist. The stem-cell and reproductive technologies described in Section 2.4, including iPSCs and in vitro gametogenesis, are being developed in part for endangered species such as the northern white rhino[10, 26], and the same methods bear on human reproductive medicine. Artificial womb research is motivated mainly by neonatal medicine, and one landmark study supported extremely premature lambs in an extra-uterine system.[31, 60] Elephants, the surrogate species for the mammoth project, carry many copies of the tumor-suppressor gene TP53 and show enhanced DNA damage responses, a finding studied as a route to understanding cancer resistance.[61, 62] Finally, the multiplex editing and cell-line validation methods used in the dire wolf process[1, 2] are general genome-engineering tools whose refinements could transfer to other applications, although we found no evidence that any de-extinction project has yet produced a medical application.

Two limits keep this from strengthening the case for de-extinction as such. First, these benefits accrue whether or not an extinct species is ever revived: the tools work on living species and cells, and the elephant cancer research uses living elephants. Second, the medical payoff is largely prospective. This section therefore supports funding the underlying technology and research base, consistent with the funding-priority argument developed in Sections 4 through 6, rather than justifying any particular revival project.

4. Case Studies

The theoretical objections raised in Section 3 do not weigh equally on every de-extinction project. A useful way to test that claim is to apply a case-by-case rubric, weighing restorative justice, animal welfare cost, ecological readiness, and resource allocation, to three projects at very different stages: the Pyrenean ibex, the dire wolf, and the woolly mammoth.[23, 24]

4.1 The Evaluative Rubric

Each project is evaluated against four criteria: restorative justice, animal welfare cost, ecological readiness, and resource allocation. The rubric is a structured tool for comparing projects, not a scoring formula. Where possible, ratings rest on documented facts, and where they cannot, on explicit ethical argument, so they should be read as the authors' reasoned judgments. Each criterion is rated as offering strong, moderate, or weak justification for proceeding with a project, where moderate indicates mixed evidence. The ratings are ordinal: "strong" means that the criterion supports going ahead, not that the project is good overall, so a project can be strong on one

criterion and weak on another. Table 1 defines each criterion.

Table 1. Definitions of the four-part rubric.

Criterion	Guiding question	Strong justification	Weak justification	Main evidence used
Restorative justice	*Did humans demonstrably cause the extinction, creating an obligation to undo it?*	*Extinction driven mainly by human exploitation or habitat destruction*	*Extinction driven mainly by natural or climatic causes, or human causation uncertain*	*Extinction-cause literature*[24, 57]
Animal welfare cost	*What documented or foreseeable harm falls on surrogates, offspring, and any founder population?*	*Healthy offspring, non-endangered surrogates, viable social structure*	*Fatal or serious defects, endangered surrogates, no social group*	*Reported birth outcomes, surrogate status, reproductive biology*[7, 22, 44]
Ecological readiness	*Do a habitat, food web, and viable population exist to receive the animal, and is any claimed ecological benefit supported by evidence?*	*Intact or restorable habitat; tested ecological claim*	*No habitat or population; benefit untested or contradicted*	*Habitat status, ecological studies*[23, 46, 48]
Resource allocation	*What is the opportunity cost relative to protecting living endangered species, and does the project yield conservation benefit beyond itself?*	*Modest cost, or clear spillover benefit to living species*	*Large capital with no spillover benefit*	*Reported funding, comparison with living-species needs*[6, 11, 54]

4.2 Pyrenean Ibex: The Baseline

The Pyrenean ibex, or bucardo, is de-extinction's founding case study, and it remains its most poignant cautionary tale. The last of its species, a female named Celia, was killed by a falling tree in 2000.[22] However, researchers had preserved her skin cells the year before. In 2003, those cells were used to create a cloned embryo, which was implanted in a domestic goat surrogate.[22] The clone was born alive, but died within minutes from a malformed lung, making the bucardo both the first extinct animal ever cloned and the only species to go extinct twice.[22, 53]

Measured against the rubric, the bucardo case scores poorly on welfare (a documented, fatal birth defect) and on ecological readiness: as a single-subspecies alpine specialist reduced to one surviving member, there was no remaining wild population to rejoin.[23, 53] It fares better on restorative justice, since centuries of human hunting pressure are generally regarded as a main driver of the subspecies' decline, which makes an obligation to undo human harm more plausible here than for the dire wolf or mammoth.[22, 53] On resource allocation, the project was small in scale, but it produced no conservation return: no population resulted, and no living species benefited. Its main value is therefore technical and precedential. It proved cross-species cloning was possible, and it remains the standard against which every later project's welfare outcomes are measured.[7, 19]

4.3 Dire Wolf: Spectacle Over Science?

The dire wolf project, detailed in Section 2.3, highlights the strong tension between productive scientific advance and commercial framing. On welfare, the picture is mixed: the pups appear to be healthy, but they are oversized compared to typical gray wolves and were gestated by domestic dog surrogates, with outcomes still being monitored rather than resolved.[2, 6] On ecological readiness, the case is weak: the animals are permanently restricted to a private preserve, and no habitat, prey base, or wild population exists to receive them.[6] On restorative justice, the case is also weak: unlike species driven to extinction by explicit human exploitation, dire wolves went extinct around 13,000 years ago as part of the broader megafauna extinction at the end of the last Ice Age, well before any human industry targeted them specifically.[24] Finally, on resource

allocation, the case is arguably the weakest of all three. Colossal, reportedly valued at over $10 billion, has directed large private capital toward a project whose headline products are three captive animals.[2, 54] Whether or not that capital would otherwise have funded conservation, the concern raised by Bennett et al. is that limited conservation attention and money spent on de-extinction may come at the expense of protecting species that still exist.[11] Beyond that, there has been significant discussion of whether the "dire wolves" are true dire wolves at all.[6, 25] As such, the project is the hardest of the three to justify as conservation spending rather than publicity.[6, 11]

The project's clearest conservation result came as a side effect rather than the headline. Using the same cloning techniques, Colossal Biosciences produced cloned red wolf pups, a critically endangered species, through non-invasive blood cloning, adding genetic diversity to a breeding population descended from a very small number of founding individuals.[1, 10] This matters for the paper's argument. The same toolkit produced a defensible conservation outcome when aimed at a living species and a contested, commercially framed outcome when aimed at an extinct one. That contrast is the empirical basis for this paper's funding-priority recommendation: review bodies should treat applications of the technology to endangered species as a distinct and more defensible category than the revival of extinct ones.[6, 10]

4.4 The Woolly Mammoth: Hope for Ecological Restoration

Of all de-extinction efforts, the woolly mammoth project has made the most ambitious claim while offering the least evidence to support it.

Unlike the bucardo, which was proof of concept, or the dire wolf, which was a demonstration, the mammoth project rests its case on ecological restoration. Colossal Biosciences argues that mammoths shaped the environment of the Pleistocene Arctic they inhabited.[55] By trampling snow cover, converting shrub and moss tundra into grassland, and raising surface albedo, large herbivores like the mammoths kept permafrost colder than it otherwise would have been.[8, 55] Reintroducing a cold-adapted proxy through de-extinction would therefore slow the rate of permafrost thaw and keep an enormous store of frozen soil carbon out of the atmosphere.[4, 55] This is the only de-extinction rationale that would justify a project on grounds other than novelty. However, the mammoth steppe was sustained by a colder, drier climate that no longer exists, and the question remains as to whether the grazing effect scales to the Arctic as a whole.[8, 48] Evidence from eastern Beringia suggests it does not; shrubs expanded despite the abundant presence of mammoths and other megafauna, with climate, rather than grazing pressure, as the primary driver of growth.[6, 48] This mechanism Colossal invokes thus scales far below the permafrost intervention the company describes.[6,48]

The technical status is equally uncertain; as of 2026, both the mammoth and its embryo have yet to exist.[56] Colossal initially projected that roughly 60 edits would convert an Asian elephant nucleus into a mammoth-like one.[56] That figure now stands at 150 and is still rising, and the target birth date has moved from 2028 to sometime in the early 2030s.[56] On welfare, the projected cost is the highest of the three cases and potentially threatens the Asian elephant species, which are themselves endangered. Using them as egg donors and surrogates subjects a threatened species to invasive procedures and a 22-month pregnancy, in an

animal whose reproductive biology is poorly understood.[7, 8, 44] Moreover, as noted in 3.3, a founder population of engineered calves would have no herd structure to be naturally raised in.[7,45]

The restorative justice claim, meanwhile, is weaker than Colossal's framing implies, as mammoths disappeared roughly 4,000 years ago due to a mix of human hunting and Holocene warming.[8, 57] If, as the argument suggests, the obligation to undo human harm applies most forcefully to the species we can be confident we destroyed, why are those species not the priority?[34, 58] The same question extends to resource allocation, with over $555 million raised directed largely at an animal that does not yet exist[54], while the living species used to produce it face a deadly herpesvirus.[59] The mammoth is therefore the strongest case for de-extinction in principle and the weakest in evidence; it is the only project whose stated purpose could answer the ecological objection raised in 3.4, and simultaneously the only one that has not yet produced an animal.[6, 10]

4.5 What the Comparison Shows

Applying the same four-part rubric to these three cases does not produce a single, unified verdict on de-extinction; it produces three distinct ones (Table 2), and that divergence is the paper's central claim. The bucardo case makes the strongest argument for caution on welfare grounds, even though its restorative justice claim is the most plausible of the three. The dire wolf case demonstrates how commercial incentives can outpace both the underlying science and the ethical justification offered to support it. The mammoth case shows that even the field's most scientifically ambitious project rests on an ecological claim that remains untested, paired with a restorative justice argument that does not clearly hold. None of this resolves whether de-extinction is defensible as a matter of principle. What it does suggest is that evaluating each project individually, rather than treating de-extinction as a single technology to be broadly endorsed or condemned, is the only approach the evidence here actually supports.

Table 2. Rubric ratings for the three case studies

Case Study	Restorative justice	Animal welfare cost	Ecological readiness	Resource allocation
Pyrenean ibex	*Moderate to Strong*	*Weak*	*Weak*	*Moderate (small scale, no conservation return)*
Dire wolf	*Weak*	*Moderate*	*Weak*	*Weak*
Woolly mammoth	*Weak to Moderate*	*Weak*	*Weak (strong claim in principle, thin evidence)*	*Weak*

5. Recommendations

The proposals surveyed in this literature diverge along one central question: should de-extinction be governed through binding, externally imposed oversight, or through industry self-regulation guided by voluntary scientific standards? This section weighs the strongest arguments on each side and ultimately defends a middle position, one that avoids both the risks of letting companies regulate themselves and the rigidity of applying a single uniform standard to a technology whose ethical stakes, as demonstrated by Section 4, differ dramatically from project to project.

5.1 Binding External Oversight

The most forceful argument for regulation comes from bioethicist Arthur Caplan, who contends that the field requires a proactive international governing body with the authority to conduct public risk assessments, establish enforceable standards, and mandate independent ethical review before any project moves forward, with especially rigorous,

non-negotiable oversight for research that touches on extinct human relatives [5]. Caplan's framework also introduces accountability measures largely missing from current practice: a liability system holding investors responsible for ecological or welfare harm resulting from a project, penalties for unauthorized attempts, and media disclosure requirements obligating companies to publish documented ethical approval rather than announce results primarily through press releases, as Colossal did with its dire wolf reveal.[1, 5]

Paganeli and Galetti offer a related but distinct proposal: a multidisciplinary review structure that pairs ecological risk assessment with biomedical-style ethics review and mandatory public engagement, overseen jointly by ethicists, conservation biologists, ecologists, bioengineers, economists, and policymakers rather than left to any single company's discretion.[6] Their proposal notably includes a funding-priority argument as well, insisting that conservation resources should be directed first toward protecting species still living before being spent to revive those already gone.[6, 11] This claim is reinforced by this paper's own finding in Section 4.3, where the dire wolf project's most concrete conservation payoff turned out to be an incidental byproduct, the cloned red wolf pups, rather than a result of the headline de-extinction effort itself.[1, 10]

5.2 Case-by-Case Evaluation

A more moderate position, advanced by philosopher Jay Odenbaugh, resists any single blanket policy and instead calls for evaluating each proposed project individually, on the grounds that welfare risk, ecological feasibility, and resource costs vary too much across candidate species to be governed by one uniform rule.[34] This mirrors, almost exactly, the conclusion this paper's own comparative case studies in Section 4 arrived at independently: the bucardo, dire wolf, and mammoth projects each succeed or fail against the same four-part evaluative framework for entirely different reasons, suggesting that a single yes-or-no policy would poorly capture the field's actual diversity.[23, 34]

Legal scholars examining the patent pathway have proposed a practical interim solution for case-by-case review: formally adopting the existing OncoMouse balancing test, which weighs documented animal suffering and environmental risk against a project's demonstrated usefulness, as a temporary standard while purpose-built de-extinction governance is developed, rather than leaving projects unreviewed until new legislation is passed.[8, 39] This approach benefits from decades of existing legal precedent, though this same test may prove more permissive toward de-extinction commercialization than most bioethicists would prefer, given that patent offices have historically defined "usefulness to mankind" broadly enough to include public engagement value alone.[8, 41]

5.3 Self-Regulation

The most liberal position, represented mostly by industry voices such as Turner and colleagues, comes in the form of a set of principles applied voluntarily by individual companies rather than through external enforcement.[10] One proposed set of principles is the IUCN's guiding principles on creating proxies of extinct species, which focuses on ecological conservation and emphasizes the nature of de-extinction technologies as proxy creators, not resurrection projects.[18] This position avoids the practical difficulty of having to build international law at a pace that matches the growth of technology.[9, 10] Though, its weaknesses have already been seen. A company marketed a genetically-modified gray wolf as a resurrected extinct species, a $10 billion project

without any external review.[6, 54] Moreover, the company continued to use the label "de-extinct" even after their chief scientist acknowledged that perfect genetic recreation of the dire wolf is impossible.[6, 25] This position is unique, being the one currently in use, and, as a result, its flaws have been well documented. Voluntary standards have not been successful in bridging the gap between marketing and scientific reality, nor have they brought the focus of de-extinction projects toward ecological preservation.[5, 6]

6. Discussion

Considering these positions along with the case studies in Section 4, this paper argues that a strong multidisciplinary body with external regulatory power, as proposed by Paganeli and colleagues, should review projects using Odenbaugh's case-by-case standard.[6, 34] This framework offers the most defensible path forward for four reasons.

First, a single universal policy, whether it allows or bans de-extinction, does not appropriately handle even the limited sample of evidence that exists today. It would treat the dire wolf, the mammoth, and the Pyrenean ibex identically despite the relevant differences in their commercial, welfare, ecological, and moral profiles, as explored in Sections 3 and 4. An external review body is necessary to enforce policy, but it must be equipped to evaluate projects individually against a consistent set of criteria, of which the rubric in Section 4.1 offers a starting point.

Second, the authority of the review body should span both a company's research and its commercialization. The dire wolf case study highlights that de-extinction can raise separate ethical complications in both scientific advancement and public communication. Colossal's gene editing succeeded, but the results may not have justified the capital and attention diverted from existing conservation efforts. Moreover, its communication was misleading in ways that affected both public opinion and policy response. A review body that incorporates both Caplan's media standards and Paganeli and colleagues' research and funding-based regulations directly addresses these issues and should be treated as the minimum requirement for ethical regulation of de-extinction projects.[5, 6]

Third, the framework should direct funding and review priority toward living endangered species. The dire wolf project's clearest conservation result was the incidental cloning of red wolves rather than the headline animals[1, 10], which suggests that the field's technology is most defensible when pointed at species that still exist. The medical benefits discussed in Section 3.7 point the same way: they stem from the underlying tools, not from reviving extinct species, and so support funding the technology rather than any particular revival. Review criteria should therefore weigh spillover benefit to living species as a positive factor and pure spectacle as a negative one.

Finally, the framework should extend Caplan's restriction on hominid research from a case-by-case judgment to a definitive limit. Section 3.6 argued that the nearly universal consensus against Neanderthal de-extinction reveals a moral boundary, protecting welfare, consent, and ecological structure, that applies to a smaller degree to every animal project.[5, 43] By making this limit explicit, this framework ensures that scientific ambition remains regulated by morality.

This framework has limits. The rubric rests on the authors' reasoned judgments across only three cases, and its ratings would need refinement as more projects and evidence emerge. The framework also leaves open who would hold jurisdiction over a technology practiced by private companies across borders, a

question Caplan's proposal for an international body addresses but does not resolve.[5] These limits argue for developing the review process in the open, not for delaying it.

7. Conclusion

De-extinction is a blanket category for a wide variety of technologies which can be used to achieve a wide variety of goals. As a result, whether de-extinction is, in principle, a valid and ethical scientific pursuit cannot be answered in a single paper. What the evidence does support is that regulation is necessary, and that it should take a specific form: binding, case-by-case, multidisciplinary external review, a firm prohibition on human-adjacent research, and a funding priority for living endangered species. The field has moved from academic theory to a billion-dollar sector faster than any institution has developed the authority to regulate it.[5, 6, 9] Closing that gap, rather than developing a general answer to the philosophical question behind de-extinction, is the relevant task for policymakers and scientists today.

8. Citations

Note on source types. † indicates news, industry, institutional, or web sources that are not peer-reviewed. ‡ indicates books, book chapters, and law reviews, which follow different review processes than journal peer review. Unmarked entries are peer-reviewed journal articles.

[1] † Colossal Biosciences. "Colossal announces World's First De-Extinction: Birth of Dire Wolves." Business Wire, 7 Apr. 2025, https://www.businesswire.com/news/home/20250407444322/en/.
[2] † Kluger, Jeffrey. "The Return of the Dire Wolf." Time, 7 Apr. 2025, https://time.com/7274542/colossal-dire-wolf/.
[3] Martinelli, Lucia, Markku Oksanen, and Helena Siipi. "De-Extinction: A Novel and Remarkable Case of Bio-Objectification." Croatian Medical Journal, vol. 55, no. 4, Aug. 2014, pp. 423–27, https://doi.org/10.3325/cmj.2014.55.423.
[4] Shapiro, Beth. "Pathways to De-Extinction: How Close Can We Get to Resurrection of an Extinct Species?" Functional Ecology, vol. 31, no. 5, 2017, pp. 996–1002, https://doi.org/10.1111/1365-2435.12705.
[5] Caplan, Arthur L. "Should Scientists Be Allowed to Bring Distant Human Ancestors Back to Life?" PLOS Biology, vol. 23, no. 9, 4 Sept. 2025, e3003384, https://doi.org/10.1371/journal.pbio.3003384.
[6] Paganeli, B., et al. "De-Extinction at a Crossroads: Ecology, Ethics, and the Future of Conservation in the Biotech Age." Ecology Letters, 2025, e70217, https://doi.org/10.1111/ele.70217.
[7] Browning, Heather. "Won't Somebody Please Think of the Mammoths? De-Extinction and Animal Welfare." Journal of Agricultural and Environmental Ethics, vol. 31, no. 6, 2018, pp. 785–803, https://doi.org/10.1007/s10806-018-9755-2.
[8] Doyle, David M., Natallia Hlushchenko, Shane McCorristine, and Kevin Healy. "Should We Patent the 'Woolly Mammoth'? Ecological and Animal Welfare Objections." Journal of Intellectual Property Law & Practice, vol. 21, no. 2, Feb. 2026, pp. 114–20, https://doi.org/10.1093/jiplp/jpaf077.
[9] ‡ Okuno, Erin. "Frankenstein's Mammoth: Anticipating the Global Legal Framework for De-Extinction." Ecology Law Quarterly, vol. 43, no. 3, 2016, pp. 581–622.
[10] Turner, Stephen D., Anna Keyte, Andrew Pask, and Beth Shapiro. "De-Extinction Technology and Its Application to Conservation." Journal of Heredity, 2025, esaf069, https://doi.org/10.1093/jhered/esaf069.
[11] Bennett, Joseph R., et al. "Spending Limited Resources on De-Extinction Could Lead to Net Biodiversity Loss." Nature Ecology & Evolution, vol. 1, no. 4, 2017, article 0053,

https://doi.org/10.1038/s41559-016-0053.
[12] † Ehrlich, Paul R., and Anne H. Ehrlich. "The Case Against De-Extinction: It's a Fascinating but Dumb Idea." Yale Environment 360, 13 Jan. 2014, https://e360.yale.edu/features/the_case_against_de-extinction_its_a_fascinating_but_dumb_idea.
[13] Green, Richard E., et al. "A Draft Sequence of the Neandertal Genome." Science, vol. 328, no. 5979, 7 May 2010, pp. 710–22, https://doi.org/10.1126/science.1188021.
[14] Novak, Ben Jacob. "De-Extinction." Genes, vol. 9, no. 11, 2018, article 548, https://doi.org/10.3390/genes9110548.
[15] ‡ Driessen, Clemens, and Jamie Lorimer. "Back-Breeding the Aurochs: The Heck Brothers, National Socialism and Imagined Geographies for Nonhuman Lebensraum." Hitler's Geographies: The Spatialities of the Third Reich, edited by Paolo Giaccaria and Claudio Minca, U of Chicago P, 2016, pp. 138–57.
[16] Harley, Eric H., et al. "The Quagga Project: Progress over 20 Years of Selective Breeding." South African Journal of Wildlife Research, vol. 39, no. 2, 2009, pp. 155–63.
[17] † The Quagga Project. "The Quagga Project Explained." Quagga Project, https://www.quaggaproject.org/the-project/. Accessed 27 Aug. 2026.
[18] † IUCN SSC. IUCN SSC Guiding Principles on Creating Proxies of Extinct Species for Conservation Benefit. Version 1.0, IUCN Species Survival Commission, 18 May 2016, https://portals.iucn.org/library/node/46248.
[19] Piña-Aguilar, Raúl E., et al. "Revival of Extinct Species Using Nuclear Transfer: Hope for the Mammoth, True for the Pyrenean Ibex, but Is It Time for 'Conservation Cloning'?" Cloning and Stem Cells, vol. 11, no. 3, 2009, pp. 341–46, https://doi.org/10.1089/clo.2009.0026.
[20] ‡ Shapiro, Beth. How to Clone a Mammoth: The Science of De-Extinction. Princeton UP, 2015.
[21] † Preston, Elizabeth. "Why 'De-Extinction' Is Impossible (But Could Work Anyway)." Quanta Magazine, 9 May 2022, https://www.quantamagazine.org/why-de-extinction-is-impossible-but-could-work-anyway-20220509/.
[22] Folch, J., et al. "First Birth of an Animal from an Extinct Subspecies (Capra pyrenaica pyrenaica) by Cloning." Theriogenology, vol. 71, no. 6, 2009, pp. 1026–34, https://doi.org/10.1016/j.theriogenology.2008.11.005.
[23] Seddon, Philip J., Axel Moehrenschlager, and John Ewen. "Reintroducing Resurrected Species: Selecting De-Extinction Candidates." Trends in Ecology & Evolution, vol. 29, no. 3, 2014, pp. 140–47, https://doi.org/10.1016/j.tree.2014.01.007.
[24] Perri, Angela R., et al. "Dire Wolves Were the Last of an Ancient New World Canid Lineage." Nature, vol. 591, 2021, pp. 87–91, https://doi.org/10.1038/s41586-020-03082-x.
[25] † "The Illusion of De-Extinction: Colossal Biosciences and the Myth of the Dire Wolf." Living with Wolves, https://www.livingwithwolves.org/the-illusion-of-de-extinction-colossal-biosciences-and-the-myth-of-the-dire-wolf/. Accessed 27 Aug. 2026.
[26] † Leibniz Institute for Zoo and Wildlife Research. "World's First Successful Embryo Transfer in Rhinos Paves the Way for Saving the Northern White Rhinos from Extinction." Leibniz-IZW, 24 Jan. 2024, https://www.izw-berlin.de/en/press-release/world-s-first-successful-embryo-transfer-in-rhinos-paves-the-way-for-saving-the-northern-white-rhinos-from-extinction.html.
[27] Lanza, Robert P., et al. "Cloning of an Endangered Species (Bos gaurus) Using Interspecies Nuclear Transfer." Cloning, vol. 2, no. 2, 2000, pp. 79–90, https://doi.org/10.1089/152045500436104.
[28] Loi, Pasqualino, et al. "Genetic Rescue of

an Endangered Mammal by Cross-Species Nuclear Transfer Using Post-Mortem Somatic Cells." Nature Biotechnology, vol. 19, no. 10, 2001, pp. 962–64, https://doi.org/10.1038/nbt1001-962.
[29] Gómez, Martha C., et al. "Birth of African Wildcat Cloned Kittens Born from Domestic Cats." Cloning and Stem Cells, vol. 6, no. 3, 2004, pp. 247–58, https://doi.org/10.1089/clo.2004.6.247.
[30] Oh, H. J., et al. "Cloning Endangered Gray Wolves (Canis lupus) from Somatic Cells Collected Postmortem." Theriogenology, vol. 70, no. 4, 2008, pp. 638–47, https://doi.org/10.1016/j.theriogenology.2008.04.032.
[31] Baron, Teresa. The Artificial Womb on Trial. Cambridge UP, 2025. Elements in Bioethics and Neuroethics, https://doi.org/10.1017/9781009544474.
[32] † "Frozen Zoo." San Diego Zoo Wildlife Alliance, https://science.sandiegozoo.org/resources/frozen-zoo. Accessed 27 Aug. 2026.
[33] † San Diego Zoo Wildlife Alliance. "San Diego Zoo Wildlife Alliance's Wildlife Biodiversity Bank Achieves Major Milestone: 11,000th Cell Line Added to Frozen Zoo." 30 Nov. 2023, https://sandiegozoowildlifealliance.org/pr/FrozenZooMilestone.
[34] Odenbaugh, Jay. "Philosophy and Ethics of De-Extinction." Cambridge Prisms: Extinction, vol. 1, 30 Jan. 2023, e7, https://doi.org/10.1017/ext.2023.4.
[35] Siipi, Helena, and Leonard Finkelman. "The Extinction and De-Extinction of Species." Philosophy & Technology, vol. 30, no. 4, 2017, pp. 427–41, https://doi.org/10.1007/s13347-016-0244-0.
[36] Genovesi, Piero, and Daniel Simberloff. "De-Extinction in Conservation: Assessing Risks of Releasing 'Resurrected' Species." Journal for Nature Conservation, vol. 56, 2020, https://doi.org/10.1016/j.jnc.2020.125838.
[37] Gunn, Alastair S. "The Restoration of Species and Natural Environments." Environmental Ethics, vol. 13, no. 4, 1991, pp. 291–310.
[38] Turner, Derek D. "Biases in the Selection of Candidate Species for De-Extinction." Ethics, Policy & Environment, vol. 20, no. 1, 2017, pp. 21–24, https://doi.org/10.1080/21550085.2017.1291831.
[39] McMahon, Aisling, and David M. Doyle. "Patentability and De-Extinct Animals in Europe: The Patented Woolly Mammoth?" Journal of Law and the Biosciences, vol. 7, no. 1, 2020, lsaa017, https://doi.org/10.1093/jlb/lsaa017.
[40] ‡ Carlin, Norman F., Ilan Wurman, and Tamara Zakim. "How to Permit Your Mammoth: Some Legal Implications of 'De-Extinction.'" Stanford Environmental Law Journal, vol. 33, no. 1, 2014, pp. 3–56.
[41] Sherkow, Jacob S., and Henry T. Greely. "What If Extinction Is Not Forever?" Science, vol. 340, no. 6128, 2013, pp. 32–33, https://doi.org/10.1126/science.1236965.
[42] Meine, Curt. "De-Extinction and the Community of Being." Hastings Center Report, vol. 47, no. S2, 2017, pp. S9–17, https://doi.org/10.1002/hast.746.
[43] Browning, Heather, and Walter Veit. "More Than Zombies: Considering the Animal Subject in De-Extinction." Ethics, Policy & Environment, vol. 25, no. 2, 2022, pp. 121–24, https://doi.org/10.1080/21550085.2022.2077916.
[44] Nicholls, Henry. "De-Extinction: A Behemoth Revised." Nature, vol. 521, 2015, p. 30.
[45] Greely, Henry T. "Is De-Extinction Special?" Hastings Center Report, vol. 47, no. S2, 2017, pp. S30–36, https://doi.org/10.1002/hast.749.
[46] McCauley, Douglas J., et al. "The Ecology

of De-Extinction." Functional Ecology, vol. 31, no. 5, 2017, pp. 1003–11, https://doi.org/10.1111/1365-2435.12868.
[47] Robert, Alexandre, et al. "De-Extinction and Evolution." Functional Ecology, vol. 31, no. 5, 2017, pp. 1021–31, https://doi.org/10.1111/1365-2435.12723.
[48] Monteath, Alistair J., et al. "Late Pleistocene Shrub Expansion Preceded Megafauna Turnover and Extinctions in Eastern Beringia." Proceedings of the National Academy of Sciences, vol. 118, no. 52, 2021, e2107977118, https://doi.org/10.1073/pnas.2107977118.
[49] Selbach, Christian, Philip J. Seddon, and Robert Poulin. "Parasites Lost: Neglecting a Crucial Element in De-Extinction." Trends in Parasitology, vol. 34, no. 1, 2018, pp. 9–11, https://doi.org/10.1016/j.pt.2017.10.007.
[50] Cohen, Shlomo. "The Ethics of De-Extinction." NanoEthics, vol. 8, no. 2, 2014, pp. 165–78, https://doi.org/10.1007/s11569-014-0201-2.
[51] Wienhues, Anna, Patrik Baard, Alfonso Donoso, and Markku Oksanen. "The Ethics of Species Extinctions." Cambridge Prisms: Extinction, vol. 1, 6 Nov. 2023, e23, https://doi.org/10.1017/ext.2023.21.
[52] Cottrell, Sariah, Jamie L. Jensen, and Steven L. Peck. "Resuscitation and Resurrection: The Ethics of Cloning Cheetahs, Mammoths, and Neanderthals." Life Sciences, Society and Policy, vol. 10, no. 3, 2014, pp. 1–17, https://doi.org/10.1186/s40504-014-0003-x.
[53] Searle, Adam. "Spectral Ecologies: De/extinction in the Pyrenees." Transactions of the Institute of British Geographers, vol. 47, no. 1, 2022, pp. 167–83, https://doi.org/10.1111/tran.12478.
[54] † Temkin, Marina. "Colossal Biosciences Reportedly in Talks to Raise New Capital at 20B–30B Valuation." TechCrunch, 20 July 2026, https://techcrunch.com/2026/07/20/colossal-biosciences-reportedly-in-talks-to-raise-new-capital-at-20b-30b-valuation/.
[55] Zimov, Sergey A. "Pleistocene Park: Return of the Mammoth's Ecosystem." Science, vol. 308, no. 5723, 2005, pp. 796–98, https://doi.org/10.1126/science.1113442.
[56] † Kluger, Jeffrey. "The Woolly Mammoth's Comeback Is Taking Longer Than Expected." Time, 24 Aug. 2026, https://time.com/article/2026/08/24/woolly-mammoth-de-extinct-colossal-biosciences/.
[57] Nogués-Bravo, David, et al. "Climate Change, Humans, and the Extinction of the Woolly Mammoth." PLOS Biology, vol. 6, no. 4, 2008, e79, https://doi.org/10.1371/journal.pbio.0060079.
[58] Sandler, Ronald. "The Ethics of Reviving Long Extinct Species." Conservation Biology, vol. 28, no. 2, 2014, pp. 354–60, https://doi.org/10.1111/cobi.12198.
[59] † World Organisation for Animal Health. "Elephant Endotheliotropic Herpesviruses (EEHV): Infection With." WOAH Technical Disease Card, 2021, https://www.woah.org/app/uploads/2021/03/elephant-endotheliotropic-herpesviruses-eehvinfection-with.pdf.
[60] Partridge, Emily A., et al. "An Extra-Uterine System to Physiologically Support the Extreme Premature Lamb." Nature Communications, vol. 8, 2017, article 15112.
[61] Abegglen, Lisa M., et al. "Potential Mechanisms for Cancer Resistance in Elephants and Comparative Cellular Response to DNA Damage in Humans." JAMA, vol. 314, no. 17, 2015, pp. 1850–60.
[62] Sulak, Michael, et al. "TP53 Copy Number Expansion Is Associated with the Evolution of Increased Body Size and an Enhanced DNA Damage Response in Elephants." eLife, vol. 5, 2016, e11994.

Journal of Ethics in Scientific and Technological Innovations

Uncaging The Truth: The Ethics of Using Animals For Research

Emily Segall[1]
[1]*Yeshiva University High School for Girls, Hollis, NY 11423*

All figures throughout the text were illustrated by Emily Segall



Abstract

The use of animals to conduct scientific research is particularly contentious with regard to its ethical considerations, and it is often misunderstood by opposers to animal research. Animals are biologically similar to humans, and are susceptible to similar health conditions and diseases as humans. Thus, they offer a reliable method of advancing in biomedical research and discovery. Those who advocate against animal research often focus on the appearance of encaged animals, and on the misperceived likelihood of harming the animals being used for research. However, in reality, there are laws protecting the animals, with an exhaustive list of practices that those overseeing and conducting animal research must abide by. Not only is animal well-being ensured when it comes to animal research, but conducting animal research can benefit both animals and humans alike – when discoveries with biomedical applications or medicinal treatments are made as a result of the animal research, they can be applied to animals as well as humans. This paper examines the ethics behind using animals for research, and reveals how the actual practices and outcomes of animal research invalidate common misperceptions regarding the practice.

1. Introduction

Animal research is the practice of studying animals in a laboratory for the purpose of biological discovery, most commonly for solving medical issues.[1] Sometimes, animals may be used for environmental and conservation research as well, such as for research aiming to protect the environment or protect a species. Animal testing specifically refers to toxicology testing, or the practice of analyzing a chemical's toxicity on animals prior to using it for human applications.[92]

The use of animals for scientific research remains a highly contentious matter. Public advocacy groups against animal research lack understanding of what animal research actually entails. These groups focus on the fact that animals are held in cages, latch onto the belief that the animals suffer as a result of being test subjects, and they argue that the outcomes of animal research do not have considerable implications for humans.[2, 3] However, these arguments are based on false misconceptions. When animals are used for research, their welfare is ensured, with several laws that researchers abide by to protect the animals involved.[4, 5]

This paper explores the practice of using animals for research, demonstrates how animal research is ethical by examining the methods of the researchers working with animals along with the protective laws surrounding the practice, and discusses biomedical advancements that have resulted from animal research.

2. A History of Using Animals For Research and the Rise of Ethical Consideration

2.1. Dating Back to Ancient Greece and the Renaissance

Animal experimentation was first adopted in ancient Greece, with the emergence

of medical studies.[6] Since dissecting humans was forbidden, the ancient Greek physicians used animals to serve as anatomical models. Such physicians would perform dissections and surgeries ("vivisections") on animals to gain physiological and anatomical understandings of living beings. Namely, Alcmaeon of Croton, Aristotle, Diocles, and Praxagoras (6th–4th century BCE) are some notable figures who conducted scientific research using animal models such as mules, embryos, pigs, monkeys, and birds for studies including anatomy and physiology.[116, 117, 118] Galen of Pergamon (2nd–3rd century CE), a Roman physician of Greek ethnicity, developed techniques for the dissection and vivisection of animals including Babary apes, pigs, sheep, and goats, and used these animal studies as a basis for his dissertations of medicine.[7]

Because the ancient Greeks had a hierarchical society, human superiority ruled out any moral concerns regarding using animals for their scientific studies.

Comparative studies between animals and humans persisted throughout the Renaissance during the 1500's, with scientific implications. Flemish anatomist Vesalius (1514–1564), a physician and surgeon, realized that many anatomical structures thought to exist in humans – due to their presence in other animals – were not, leading him to break civil and religious rules and illegally dissect human cadavers.[119] Vesalius then published highly accurate descriptions of the human anatomy, challenging the authority of the classical authors. He would also examine the similarities and differences between the internal structure of humans and other animals, setting the foundations of modern comparative anatomy.

2.2. Animal Studies During the Enlightenment

Throughout the Age of Enlightenment, animal experimentation continued to be used as a means of gaining physiological knowledge and understanding.[6] At this point in time, animal sentience – the capacity of an animal to have feelings with a subjective character, including pleasure, pain, hunger, curiosity, loneliness, and fear – became recognized.[120] Animal sentience greatly contributes to ethical debates, as sentient animals are capable of suffering, providing a necessity for moral consideration in animal research. It is also therefore relevant for policies seeking to protect animal welfare. The concept of animal research gained opposition when animal sentience became recognized due to the moral concerns regarding the possibility of animals to suffer. Still, scientists found ways to justify using animals for their studies, such as with René Descartes' (1596–1650) statement of animals being "machine-like" and lacking intelligence.[8] Other philosophers, however, such as John Locke, argued that animals have sentience and should not be used for experimentation.[9] These contradictory arguments demonstrate the contention over using animals for research, even during historic times. The emergence of acknowledging animal sentience did not become particularly noteworthy during the early time periods, though, and it was not until later centuries that the ethics behind animal experimentation were taken into deeper consideration.

Animal studies fostered significant developments in biological understanding. Physician William Harvey (1578–1657) published largely accurate descriptions of blood circulation and heart function, providing another such example of early advancements that stemmed from animal research.

2.3. The Rise of Ethical Concerns

The moral perspective of using animals for scientific experimentation re-emerged in the eighteenth century, and opposition to experimenting on animals rose across the globe. Philosophers such as Voltaire (1694–1778),

Jean-Jacques Rousseau (1712–1778), and Jeremy Bentham (1748–1832) recognized animal sentience as a factor to be considered, arguing that if animals can feel suffering and pain (sentience), they should not be used for experimentation.[6]

By the nineteenth century, anti-vivisection sentiments rose.[10] As the world saw an increase in laboratory experimentation to shape medical sciences, the public became more aware of the practice of vivisection, leading to a surge in opposition to the practice. In 1822, Richard Martin's (1754–1834) Cruel and Improper Treatment of Cattle Act (3 George IV. c. 71) became a law in the UK, making it illegal to harm cattle.[10] Prior to this Act, it had only been illegal to inflict harm upon someone else's animal, not one's own. Martin's Act did not mention vivisection, likely because he was unaware of it, which could be due to the fact that it was not particularly widespread in Britain. The British public became more aware of animal experimentation when French physiologist François Magendie (1783–1855) performed a demonstrative experiment at London's Windmill Street anatomy school, prompting public commotion from a moral standpoint.[10] Martin then argued against vivisection as well, though vivisection was not considered a cruel, aimless act, so it was not covered as illegal under Martin's Act. The rise of moral concern over animal abuse sparked the founding of the Society for the Prevention of Cruelty to Animals (SPCA), in 1824, highlighting society's proactive attempts to suppress immoral activity towards animals. After enough public outcry, the Cruelty to Animals Act (the "Vivisection Act") was passed in Britain in 1876.[11]

In the late nineteenth and early twentieth centuries, the American Medical Association (AMA) played a key role in creating guidelines for ethical conduct in animal experimentation for scientific and medical purposes.[12] The AMA favored vivisection for the sake of advancing in medicine and protecting human health.

Ethical concerns over animal research remain prominent to this day. Certain sectors of society argue that using animals for scientific purposes is immoral, fearing that the animals will suffer in the process, while others – largely those who understand the moral circumstances under which animal research is conducted – recognize the greater good that it provides to both humans and animals alike.[13, 14, 15]

3. Biological Comparisons: Which Species, and When?

Humans and animals are biologically very similar, are naturally susceptible to similar diseases and conditions, and share much of the same DNA. Non-human mammals are especially similar to humans, sharing between 50% and 99% of the same DNA, depending on the species.[16] Thus, they offer a reliable means of biological study.

3.1. Rodents

Rodents are readily available, reproduce quickly and easily, and are small, making them easy to maintain and take care of.[22] As a result of these factors, rodents are optimal for research.[21, 22]

Mice (namely the domestic North American and European mouse *Mus musculus*) and rats (namely the Norway rat, *Rattus norvegicus*) are most commonly used as animal models for research.[17, 21] Approximately 95% of all laboratory animal models are mice and rats.[18, 19, 22] In fact, rats, mice, and humans share approximately 95% of their 30,000 genes.[20] Mice and rats tend to be used in research to study the roles of genes in both normal biological functions and disease-related processes, as well as for research on cognition, aging, memory, and behavioral studies.[17, 21] These animals are also used to observe the effects of certain treatments and drugs, and

assess the likelihood of these treatments being beneficial or harmful to humans.[17, 21]

Hamsters are also used in research.[21] As they are mostly free of diseases and susceptible to induced viruses, hamsters are an ideal model for studying various diseases and viruses. Hamsters are used to study conditions including obesity, carcinogenesis, prostatic disease, toxicoses, infectious diseases and slow viruses, dental cavities, bronchitis, and teratogenesis.[21]

Guinea pigs serve as animal models most often used for studying embryogenesis, and advancing preimplantation development and stem cell research.[26] This is largely because guinea pigs' reproductive systems closely resemble that of humans, even more so than other animal species.[27] In addition, guinea pigs have similar airway physiology to that of humans, so they are used in research on respiratory virology and tuberculosis as well.[21]

Gerbils (namely the Mongolian gerbil, *Meriones unguiculatus*) have anatomical variations in the blood supply to the brain, and are therefore most commonly used in studies of cerebral ischemia or stroke.[22, 44]

Finally, chinchillas (long-tailed chinchilla, *Chinchilla lanigera*, and the short-tailed chinchilla, *Chinchilla chinchilla*) have remarkably large inner ear anatomies, making them preferred rodents for auditory studies and otitis media.[22]

3.2. Rabbits

While rabbits are not the most common in animal research, they serve as strong models for studying immunity and antibodies.[23] For example, European *Oryctolagus cuniculus* rabbits have been used to study vaccine development.[24, 25] Some studies on genetic comparisons between humans, rabbits, and mice suggest that European *Oryctolagus cuniculus* rabbits may even offer a stronger model for studying innate immune systems than the mouse, considering their lower genetic diversity to humans compared to that of mice.[29] Moreover, rabbits are easy to breed, so they do not have to undergo invasive procedures.[23]

One limitation of using rabbits for research, however, is that, despite their docility and positive responses in the laboratory environment, they are large and require mental and social stimulation, making them less ideal for easy housing.[23, 28] This is also why mice tend to be more practical models for research.[29]

Rabbits' larger size can serve as an advantage, though, when conducting research involving biomedical device implantations – it is far more practical to perform surgical implantations of such devices on larger rabbits than on comparatively smaller rodents.[22] Furthermore, their larger size makes rabbits highly suitable in polyclonal antibody production studies, since they have a higher volume of blood than rodents.[30]

3.3. Zebrafish

Zebrafish (*Danio rerio*) have a history of being used for research on embryonic development, and are increasingly expanding into studies related to biochemistry, molecular biology, cell biology, neurological sciences, and genetics.[17, 22]

Around 70% of human genes have at least one clear zebrafish ortholog (genes separated by speciation).[31, 121] This genetic similarity to humans in conjunction with their small size, large numbers, and frequent spawning make zebrafish strong research models.[17, 22]

3.4. Amphibians and Reptiles

Amphibians, such as frogs and salamanders, along with reptiles, including lizards and snakes, are most often used for studying ecology and evolution.[22] Amphibians play key roles in research on embryonic development, metamorphosis, regeneration, physiology and thermoregulation, and the effects

of climate change on biological processes.[17, 22, 32, 33, 34] In contrast, reptiles tend to be used for cardiovascular studies – due to their simplistic cardiovascular systems – in addition to research on immune response mechanisms, endocrine systems, and unique reproduction methods.[17, 22]

The *Xenopus tropicalis* frog, which is commonly used in animal research involving amphibians, has over 1,700 genes that are notably similar to genes in humans related to conditions like cancer, asthma, and heart disease.[35] Additionally, genes in frogs have similar neighboring genes to humans nearly 90% of the time.[35] With these genetic similarities, such animals provide reliable models for research.

Because amphibians and reptiles are ectotherms, meaning they are unable to regulate body temperatures above that of the surrounding environment internally, they are used in studies on environmental endocrine disruptors – natural or man-made substances that interfere with the body's normal hormone production, release, transport, metabolism, or elimination – on these animals.[22, 32, 33, 34] As ectotherms, amphibians' physiological processes, developmental pathways, and reproductive systems are intimately linked to and controlled by their immediate surroundings. This, as well as that amphibians live in both aquatic and terrestrial environments, makes amphibians ideal for studying the effects of endocrine disruptors on animals in various environments. Amphibians and reptiles also model for research on ecological issues such as climate change, for the same reason.[33]

Reptiles have unique cardiovascular systems, consisting of three chambers: two atria and one shared ventricle (unlike how humans and other mammals have two atria and two ventricles).[36] As a result, they offer a means of studying diverse cardiovascular systems.[37]

3.5. Birds

Birds have been used in a wide range of research topics, from aging and cognition to parasitology and infectious disease studies.[22] Genetic similarities between human and bird species have been identified, suggesting that birds are another optimal animal choice for research.[22, 43]

Responses to pathogens in birds (avians) and mammals are very similar, as both exhibit the similar innate and adaptive immune responses.[38] But, when examining the organs, cells and molecules of the immune response in mammals and birds on a more detailed scale, there are noticeable differences between mammalian and avian immune response mechanisms.[38] In essence, the particular haplotype (a physical grouping of genomic variants that tend to be inherited together) of the B-F/B-L region of the chicken B locus – the group of genes responsible for determining life and death in response to certain infectious pathogens and vaccines – is much smaller and simpler in birds than the typical mammalian major histocompatibility complex (MHC).[39, 122] This makes birds ideal for research because their single-gene expression eliminates genetic redundancy, allowing scientists to directly link a specific gene to disease resistance or susceptibility.[123]

Chickens (*Gallus domesticus*) are the most common bird species studied in biomedical and agricultural research.[22] Chickens tend to be used to study immunology in particular, and chicken eggs enable the production of many vaccines.[17, 38, 42] Chicken embryos are also used in toxicology studies, as they are cost efficient, easily accessible, and exhibit biological similarities to humans and other vertebrates, especially during early developmental stages.[40, 41]



Figure 2. Drawing of chicken eggs.

Quail (*Coturnix coturnix and Coturnix japonica*) are used in research on reproductive biology, as they readily mate and exhibit sexual behavior in captivity.[22]

Parrots of the genus *Amazona* (Amazon parrots) and budgies (*Melopsittacus undulates*) are used as research models for cognitive studies, aging, behavioral and sensory studies, as well as for veterinary medicine and diagnostics.[22] The African gray parrot (*Psittacus erithacus*) is used in research on cognition and communication, too.[22] Also used for cognitive and neurobiological studies are zebra finches (*Taeniopygia guttata*) and other small finches, along with canaries (*Serinus canaria*). These small creatures are relatively easy to house due to their small sizes, and they portray ideal characteristics for research models with their compatibility in groups and easy breeding.[22]

Pigeons (*Columba livia*) are used in research on atherosclerosis, plus neurological studies including comparative psychology, neuroanatomy, and neuroendocrinology, and they are studied to gain stronger understandings of their navigational skills and memory.[22]

Barn owls (*Tyto alba*) are another bird species used in animal research, as nocturnal models for neuroanatomy, vision, and auditory related research.[22]

3.6. Swine

Domestic swine (*Sus scrofa domesticus*) are closely related to humans when analyzing anatomy, genetics, and physiology, with similar organs in terms of functional features and size.[46] These models are widely used as a result of their favorable reproductive capacity with short gestation periods and large offspring. Swine are useful models for research on coronary artery disease and other cardiovascular conditions like congenital heart disease and hypertension, as well as for studies on diabetes, hemophilia, and obesity.[21, 45] Swine are often used in xenotransplantation – grafting or transplanting organs or tissues between members of different species – as they have similar physiology and organ sizes to humans.[124] And, due to their skin's similar resemblance to human skin, they are common skin models.[46, 47]

3.7. Dogs, Cats, and Nonhuman Primates

Dogs, cats, and nonhuman primates are extremely uncommon in animal research, accounting for less than 1% of all animal models.[48]

In the rare instances in which these animals are used, it tends to be for biomedical research. Dogs' size, trainability, and physiology allow them to be suitable animal models in research studies on prosthetics, orthopedics, and endodontic therapies, as well as for translational research including chronic implantations and telemetry.[21, 49] Purpose-bred beagles tend to be the preferred dog model, for their docile temperament and uniform size.[21]

The use of cats in research has significantly declined. In the past, cats have been used to study both inherited and infectious diseases, plus epilepsy and spinal cord injuries.[21, 50]

Nonhuman primates (NHPs), such as chimpanzees (*Pan troglodytes*) and bonobos (*Pan paniscus*), are the closest relatives humans have, sharing around 98.8% of their DNA with humans.[51] However, as they are considered endangered species, many NHPs have been prohibited from being used for biomedical research in a number of countries, including the

US.[21, 53, 54] Still, NHPs have led to major advancements in biomedical research, especially when studying age-related neurodegenerative diseases, such as Alzheimer's and Parkinson's, and infectious diseases, such as HIV/AIDS.[21, 55] NHPs have also been used to assess the safety and efficacy of treatments and vaccines, like immune-suppressing drugs for organ transplantations and the COVID-19 vaccine.[17, 21]

3.8. Alternatives to Animal Models

Scientists aim to minimize the number of animal models used for research, and whenever the researchers can achieve the same research goal using a non-animal model, they opt to do so.[74, 48]

Alternatives to animal models include *in vitro* and *in silico* models. *In vitro* ("in glass") refers to processes that can occur outside of the living cell, such as tissue or organ culture.[72] *In silico* refers to using computers, or computer simulations, to perform scientific experiments.[73] Both *in vitro* and *in silico* methods have been used in past research to make scientific and biomedical advancements, one example being for toxicological studies.[72, 73]

Despite their continued uses in research, *in vitro* and *in silico* models have their limitations. When performing *in vitro* experiments, it can be difficult to conduct completely accurate, detailed studies, as a single cell grown in a petri dish behaves differently than cells in a multicellular environment, as a part of a complex organ system.[74] Similarly when conducting *in silico* studies, although using organoids or AI models can be incredibly beneficial in some ways, it can be disadvantageous in the sense that it is crucial to study biological systems in their complex, natural state, not solely *in silico.*[74]

Currently, the alternative research models available can supplement but not completely replace *in vivo* experiments, and animal research therefore remains crucial to scientific and biomedical discovery.[74, 78]

4. The Laws Behind Animal Research

Modern animal research requires highly specific conditions and care for the animals involved, and extensive regulatory oversight.[55, 56, 57]

In the United States, there are federal agencies such as the US Department of Agriculture (USDA) and the National Institutes of Health's Office of Laboratory Animal Welfare, which enforce the Animal Welfare Act (AWA), Animal Welfare Act Regulations (AWAR), and the Public Health Service Policy on Humane Care and Use of Laboratory Animals (PHS Policy), as well as the Institutional Animal Care and Use Committees (IACUC).[57, 58, 59, 69] Furthermore, the Association for Assessment and Accreditation of Laboratory Animal Care International (AAALAC International) provides independent accreditation for animal care programs in industry, academia, and government.[69] These agencies and laws provide and oversee regulations and frameworks for the safe-handling of animals in research, to protect the animals being used in such studies.

The AWA is a Federal Law that regulates the treatment of animals in research, as well as in teaching, testing, exhibition, transport, and by dealers.[58] Some of what the AWA establishes includes the minimum standards of care and treatment for animals, and requires an attending veterinarian with appropriate authority on site as an overseer to ensure that adequate care for these animals is provided.[126] In addition, it provides specifications for the proper care of each animal species being used for research. The AWA has been updated multiple times since its initial establishment to achieve the regulations it provides today.[58]

The PHS Policy provides the regulatory framework for the use of live, vertebrate animals in any activity supported or conducted by the PHS agencies and U.S. Department of Health and Human Services components, or with entities that have a memorandum of understanding (MOU) with National Institutes of Health (NIH) to apply the PHS Policy. The PHS is designed to supplement the "U.S. Government Principles for the Interagency Research Animal Committee's "Utilization and Care of Vertebrate Animals Used in Testing, Research, and Training," and it requires compliance with the AWA. It also requires a written Assurance – thoroughly describing the institution's program for the care and use of animals in the facilities – acceptable to the PHS, the Office of Laboratory Animal Welfare (OLAW), Office of the Director, National Institutes of Health (NIH), and signed by the Institutional Official.[127] While the AWA excludes purpose-bred rats, mice, and birds, however, the PHS Policy covers all live vertebrates – ensuring that such animals are always under adequate provision and oversight.[128] Much like the AWA, the PHS Policy has been updated a number of times since its establishment in the early 20th Century to achieve the modern standards it provides.

The IACUC must approve of the vertebrate animals being used for research prior to conducting any particular experiment.[68] They review circumstances including the value of the animal research to human or animal health, the design of a particular study, the rationale and justification for animal use in the experiment, the choice of species, alternatives to animal use, animal care procedures, and how the welfare and well-being of the animals will be ensured throughout the study.[68]

Moreover, the UK-based National Centre for the Replacement, Refinement and Reduction of Animals in Research (NC3Rs) works to minimize and improve the use of animals in scientific testing and research, and has developed the Animal Research: Reporting of In Vivo Experiments (ARRIVE) guidelines to improve the transparency, reliability, and reproducibility of animal research.[70, 71]

4.1. The 4 R's

Added to the initial "3 R's" proposed in the 1900's, standing for "Reduction, Refinement, and Replacement," is "Responsibility." These principles illustrate the ethical practices in animal research.[60, 61, 62, 65]

Reduction. The first "R" stands for "Reduction." Meaning, the experimental design is examined to ensure that researchers have reduced the number of animals being used in a research study to the minimum required for reliable data.[60]

Refinement. The second "R" stands for "Refinement," and it encompasses improvements in laboratory procedure that minimize the harmful effects of the proposed experiments on the animals.[60] This is implemented to reduce the pain, distress, and suffering to as minimal as possible – ideally so that the animals are susceptible to none of these conditions – to seek improvement in general animal welfare.[60] Examples of Refinement include, but are not limited to, improved living conditions for research animals, proper training of those handling the animals, application of anesthesia and analgesia when necessary as well as the need for euthanasia of the animals at the end of the experiment to limit the animals' suffering.[60]

Replacement. "Replacement" refers to methods to replace or avoid the use of animals in experiments, such as in silico methods, or computerized techniques, and in vitro methods, such as cell and tissue culture.[60] When there is a possible alternative to animal models for a specific study, that alternative is pursued. Furthermore, there are relative replacement methods by using invertebrates like nematode

worms (including *C. elegans*), fruit flies (*Drosophila melanogaster*) and other microorganisms instead of vertebrates and higher order (more complex organisms, like mammals) animals.[17, 60, 129]

Responsibility. Expanded from the original "3R's" is an additional, fourth measure of "Responsibility," to further protect animal welfare and cover any gaps or limitations that may exist within the initial three principles.[60, 63] "Responsibility" encompasses improvements in animal well-being during animal experimentation, and it accounts for researchers' empathy towards the animals.[60, 63] It includes acknowledging animal sentience and consciousness, promoting social interactions within the animals' environments, as well as providing animals with proper stimulation and enrichment when being held in the research environment.[60, 63, 64]

4.2. Extensions to the 4 "R" Principles

Taking these established principles even further and expressing greater concern for animal welfare leads to the "3 S's." the "3 V's," the "4 F's," and the "6 P's".[66]

The "3 S's" – standing for "Good Science," "Good Sense," and "Good Sensibilities" – are designed to ensure that (1) the goal of research is to do "good science" and contribute new knowledge; (2) the researchers use "good sense" and ensure that the proper model, whether *in vitro*, *in vivo*, or *in silico*, is used for any particular experiment; and (3) the researchers implement "good sensibilities" in that they understand and express proper care towards the animals involved.[66]

The "3 V's" highlight the validation of animal models, using "Construct Validity," "Internal Validity," and "External Validity".[66] The first "V" of "Construct Validity" is geared towards the question of whether the animal model is valid for the specific scientific objective of an experiment.[66] The second and third "V's" of "Internal Validity" and "External Validity" respond to the questions of whether the model approach and the protocol are designed appropriately, and how likely it is that the research will broaden to other animals or humans based on its goal.[66]

Next, the "4 F's" propose "Fundamental Principles" to further emphasize the responsibilities of the researchers.[66] The first Fundamental Principle, the biomedical research principle, states, "research is among the most noble and imperative of human endeavors, relying on the scientific method to prevent, alleviate, and cure the pain, suffering, distress, fear, anxiety, disability, infirmity, and death associated with disease".[66] The second Fundamental Principle, the "animal research principle," states, "the development of knowledge necessary for the improvement of the health and well-being of humans as well as other animals may well require *in vivo* experimentation with nonhuman animals".[66] The third Fundamental Principle, the principle of medical research on human subjects, states that such research "should be so designed and based on the results of animal experimentation and a knowledge of the natural history of the disease or other problem under study that the anticipated results justify the performance of the experiment".[66] Finally, the fourth Fundamental Principle establishes that "animal research should aim to minimize pain or distress to the animals in the achievement of the scientific goals of the research".[66]

Lastly, the "6 P's" are additional principles that balance social benefit with animal welfare, presenting an appropriate moral framework to address these two aspects of animal research.[66, 67] These six principles are divided into categories of the "Principles of Social Benefit" and the "Principles of Animal Welfare," each containing three principles.[67]

The first Principle of Social Benefit is "The Principle of No Alternative Method," stating, "Use of animal subjects must be the sole ethically acceptable way to address a research problem whose solution offers the prospect of social benefit".[67] The second Principle of Social Benefit is "The Principle of Expected Net Benefit," which states, "The prospect of social benefit from a research study must outweigh its expected costs and risks to human beings".[67] The third Principle of Social Benefit, "The Principle of Sufficient Value to Justify Harm," states that "The prospect of a net benefit for human society from a research study must be sufficiently valuable to justify expected harms to animal subjects".[67]

The three Principles of Animal Welfare are similarly constructed to ensure morality in animal research. The first Principle of Animal Welfare is "The Principle of No Unnecessary Harm," stating that "Animal subjects must not be harmed unless a particular harm is necessary for and morally justified by scientific purposes".[67] The second Principle of Animal Welfare is "The Principle of Basic Needs," protecting the animal needs that are specified and enforced by the aforementioned organizations and policies, stating that "Animal subjects' basic needs must be met in the conduct of studies unless failure to meet specific basic needs is necessary for and morally justified by scientific purposes".[67] This includes but is not limited to nutritional, shelter, stimulatory, and veterinary needs.[66] The last, sixth principle is "The Principle of Upper Limits to Harm," stating, "Animal subjects must not be caused to endure severe suffering for a lengthy period of time".[67] It further notes, "In rare, extraordinary cases, exceptions may be warranted if the research is necessary for and morally justified by critically important social and scientific purposes".[67]

4.3. Euthanasia

Euthanasia is defined as "the act of inducing humane death in an animal by a method that induces rapid loss of consciousness and death with a minimum of pain, discomfort or distress".[82] Specific guidelines for the humane euthanasia of animals have been set, tailored to each species, to ensure it is done properly and ethically.[82, 83] Euthanasia may be performed at the end of a study if further research needs to be conducted using the experimental animal's organs, or if the experimental animal is unfit for adoption.[80, 83-85] When adoption of laboratory animals is a feasible, safe option, facilities may create adoption programs as opposed to euthanizing the animals.[88]

Different methods of euthanasia may be performed depending on the circumstance and type of animal. Common methods include inhalation of anesthesia gas, typically for rodents and other small animals, as part of a two-step process with a secondary physical method of euthanasia like decapitation or cervical dislocation; inhalation of CO_2, which comes with its own unique set of guidelines; immersion agents, typically for aquatic species, and often paired with a secondary method; cervical dislocation, primarily for small birds, mice and immature rats, and requires additional training and is performed under anesthesia unless specifically approved by the IACUC; decapitation, for rodents and small rabbits, with additional training and anesthesia required; injectable barbiturate agents; and exsanguination/cardiac perfusion, which is acceptable with conditions, requires animals to be anesthetized, and the IACUC strongly encourages using injectable anesthetic agents prior to such methods.[133] There is also an assessment criteria for the confirmation of death in euthanized animals, as well as special considerations for certain animals such as rodents, since they are particularly resistant to euthanasia by inhalation.[133] The American

Veterinary Medical Association (AVMA) provides an elaborate set of guidelines for the safe and proper euthanasia of animals, when euthanasia is necessary.

When deciding on whether or not to perform euthanasia on a particular animal, veterinarians analyze the animal's welfare or quality of life.[83] Scientists have established the accepted definition of "welfare," in this case, as having three components: that the animal functions well, feels well, and has the capacity to perform innate or species-specific adaptation behaviors.[83] The decision on what to do with the animal(s) involved at the end of each study is made by evaluating these circumstances, and determining if the animal(s) can continue to live comfortably. If euthanizing a particular animal will result in the least amount of suffering for that animal, euthanasia is performed; if the animal still has an optimal quality of life, euthanasia is avoided.

Animals may also be euthanized to further study their organs, tissues, and/or blood, if such parts need to be obtained in a way that would not allow the animal to survive.[134, 135] Still, when it is possible to do so, anesthetics are used to allow for both continued experimentation and for the animal to survive.[134]

Overall, animal research is subject to regulations and guidelines to protect the experimental animals, and the researchers and veterinarians involved in animal research follow specific guidelines to properly care for the animals, ensuring the applications of ethical and moral circumstances during the research. Some examples of facilities that comply with such regulations include, but are not limited to, the Harvard Medical School, which falls under the IACUC as well as adheres to the AWA, PHS, AAALAC International, plus the Massachusetts Department of Public Health, Massachusetts Division of Fisheries and Wildlife, and the U.S. and Massachusetts Drug Enforcement Administrations; the University of Missouri, overseen by the Office of Animal Resources (OAR), the Animal Care and Use Committee (ACUC), and the Animal Care and Quality Assurance office (ACQA); and Noble Life Sciences, which is accredited by the AAALAC, and has FDA inspected Good Laboratory Practice (GLP)-compliant facilities and operate in full compliance with the AWA, the PHS Policy on Humane Care and Use of Laboratory Animals, and the NRC Guide for the Care and Use of Laboratory Animals.[57, 131, 132] These facilities exemplify how laboratories carefully follow the laws and regulations that have been set, and such regulations do not go unadhered to.

5. Modern-Day Moral Concerns and Misconceptions

5.1. "Animal Suffering is Inherent to Animal Research"

When conducting animal research, animal welfare comes first and foremost.[48, 75] Some people mistakenly believe that animals are held isolated in captivity, and forced to endure pain and suffering through induced diseases and injuries.[76, 77] These people may latch onto the belief that experimental animals lack enrichment and are deprived of their natural necessities.[76, 77] However, as previously discussed, such statements are false. Animals being used for research are well-cared for and protected under established laws and regulations ensuring they are provided with adequate shelter, nutrients, and stimulation.[48, 55-71, 75] Animal research is also heavily monitored and carefully overseen from beginning to end, ensuring the safety and welfare of all animals involved.[75]

5.2. "Animal Research Does Not Reliably Predict Human Outcomes"

Opponents of animal research may claim that due to the differences between humans and animals, it is too difficult to apply experimental data from animal research to

human conditions.[76] They may also claim that drugs which may appear effective in animals fail in humans more often than they exhibit positive outcomes.[76] While research may not have a guaranteed 100% success rate, animal research has led to a wide array of pivotal scientific and biomedical advancements for both humans and animals alike.[78-81] Even though humans and animals have some biological differences, they are largely similar and share much of the same DNA.[17, 19-54, 75]

5.3. Euthanasia Concerns

Some may express concern over euthanasia being inhumane and unnecessary, with the assumption that it is the most preferred method to conclude a study.[86, 87] However, the guidelines for the humane euthanasia of animals and the use of adoption programs when possible to avoid euthanizing the animals demonstrates the ethical practices in animal research and the protection of animal safety, overriding these fears.[88]

5.4. Sourcing Laboratory Animals

A misconception regarding the sourcing of laboratory animals is that they come from random sources, including domesticated pets.[48] This sparks emotional outrage, and instills fear in people, especially in those who believe the misconception that animals used for research are subject to pain and suffering.[130] However, laboratory animals are purposefully bred by licensed, USDA-approved breeders specifically for research use, and animals are not taken from homes or the wild to be taken into research facilities.[48, 89-91]

5.5. Animal Testing: Cosmetics

The terms *animal testing* and *animal research* differ in that animal *testing* most commonly refers to toxicology studies, whereas animal *research* refers largely to biomedical studies.[92] Using animals for cosmetic testing does not mean that cosmetics are applied to animals to evaluate a product's safety. Rather, it refers to the toxicity tests that are performed using the chemicals found in certain products.[92] There are a variety of types of toxicology testing, including acute toxicity, eye and skin irritation, repeated-dose chronic toxicity, carcinogenicity, developmental and reproductivity, neurotoxicity, mutagenicity, and biological screening tests.[92] Concerns surrounding animal testing include misconceptions of the experimental animals not receiving pain relief; however, when animals are used for scientific experimentation, they are well cared for and receive analgesics and anesthetics.[60, 95, 98]

Acute toxicity tests consist of single doses at concentrations high enough to exhibit toxic effects, and are often used to evaluate the relative toxicity of certain substances.[92] The Lethal Dose 50% (LD_{50}), the dose of a test substance at which half the test animals can be expected to die, is one example of this, but it is, for the most part, phased out and only conducted under circumstances in which alternative tests cannot be used.[94, 97]

Eye and skin irritation tests, usually consisting of a single exposure, are used to predict the toxicity of accidental exposure, as well as to develop warnings for product handling.[93] For instance, the Draize test has been used to test eye irritation, in which a test substance is applied to one eye of an adult rabbit under analgesia and anesthesia, with the untreated eye serving as a control.[95] Low-volume eye-irritation tests (LVET) have been developed as a more ethical alternative to Draize testing, using lower volumes of the substance being tested, and using no forced eyelid closure.[95] Such alternatives provide more moral methods to animal tests that have been used in the past.

Repeated-dose chronic toxicity tests, most commonly performed on rats, involve repeated exposures to substances for time

periods ranging from multiple weeks to a year, to determine possible effects of long-term exposure of a particular substance.[93, 96] Similarly, carcinogenicity tests involve repeated exposures to substances for most of an experimental animal's lifespan to discover human carcinogens.[93]

Developmental and reproductive toxicity tests utilize a variety of procedures – usually performed on rats and rabbits – to detect the potential of substances to cause infertility, miscarriages, and birth defects.[93]

Neurotoxicity tests are performed using various doses and exposures to observe toxic effects of a substance on the nervous system, such as behavioral changes, lack of coordination, motor disorders, and learning disabilities.[93]

Mutagenicity tests are used to determine whether genetic material of germ or somatic cells has been changed due to the substance being tested.[93]

Lastly, biological screening tests examine the biological activity of organic compounds.[93]

Whenever animals are used in a laboratory setting – whether for the purpose of biomedical research or toxicology testing – the researchers and those overseeing the experiments conducted (such as veterinarians) comply with the laws and regulations surrounding the use of animals for experimentation, as previously discussed in Section 4.

6. Positive Outcomes of Using Animals for Research

Animal research has led to a plethora of scientific and biomedical discoveries for both humans and animals. The following are some briefly described examples of how animal research has significantly contributed to advancements in science and medicine.

6.1. Smallpox

Research involving purifying extracted cowpox material from calves led to the creation of the smallpox vaccine, creating the first vaccine and leading to vaccinations against measles and chickenpox.[114]

6.2. Polio Vaccine

Poliovirus (polio) is a highly contagious viral disease involving the spinal cord and brain, severely affecting those infected, disabling breathing without a ventilator or "iron lung".[104, 105] By studying polio in mice and rhesus monkeys, the polio vaccine was developed, eliminating wild poliovirus in the US and Europe, and saving millions of lives.[104-106]

6.3. Cancer Immunotherapies

Thanks to animal research, cancer therapies have been developed, helping dogs, as well as millions of humans.[99-103]

6.4. HIV/AIDS Breakthroughs

Scientists are studying HIV/AIDS in various animal models including mice, NHPs, and felines, leading to major advancements in treating these viruses.[106-111]

Feline immunodeficiency virus (FIV) is a lentivirus (a genus of retroviruses that cause slowly progressive often fatal human and animal diseases) that causes Acquired Immune Deficiency Syndrome (AIDS) in domestic cats, much like how Human Immunodeficiency Virus (HIV) causes AIDS in humans.[106, 107, 125] Similar viruses such as Simian Immunodeficiency Virus (SIV) can be found in NHPs as well.[108-111] Studies of these animals have helped develop antiretroviral therapy (ART), a combination of multiple drugs that can be taken to reduce the amount of HIV in one's body.[110]

6.5. Neurodegenerative Diseases and Psychiatric Disorders

Examples of common neurodegenerative diseases are Alzheimer's, Parkinson's, Huntington's, muscular dystrophy, and multiple sclerosis, which can affect balance, movement, talking, breathing and heart function.[112]

Mice and NHPs have provided scientists with knowledge of the workings of the brain, which can help with all neurological conditions, including Alzheimer's and autism, along with psychiatric conditions like bipolar disorder.[114]

6.6. Flu and COVID-19 Vaccines

Ferrets have been important models for studying the influenza (flu) virus, as they are naturally susceptible to flu variants A and B, with such infections progressing similarly in ferrets as they do in humans.[114]

Additionally, the most common way that the flu vaccine is made is from chicken eggs.[42] And even when it is made using alternative methods, mammalian cells are used for its development.[42]

Moreover, prior to clinical trials, the COVID-19 vaccine needed to be tested in pre-clinical studies using animal models.[115] Mice, hamsters, ferrets, NHPs, and pigs served as research models in developing the COVID-19 vaccine.[115]



Figure 3. Representation of a vaccine.

6.7. Diabetes

Monitoring blood sugar levels in rabbits and mice, researchers have been able to use insulin in dogs, and later in humans.[114] Furthermore, researchers have been developing an insulin shot that would last a whole month – as opposed to the weekly, or even daily, shot that type 2 diabetics must take – which has so far demonstrated efficacy in mice and NHPs.[114] In such a way, animal research has improved the lives of diabetics, which not only includes humans with diabetes, but diabetic animals, too.

These examples are only a handful of the many achievements that have stemmed from animal research. Animal research has played a crucial role in science and medicine, enhancing millions of lives across the globe.

7. Conclusion

Animal research is vital for scientific and biomedical discovery. The laws and regulations that are employed during animal research protect the animals involved, ensuring ethical practice. Animal research has led to major advancements, including numerous treatments and fields of knowledge that are available today and continue to be explored.

8. References

[1]"What Is Animal Research?" Understanding Animal Research, www.understandinganimalresearch.org.uk/animal-research/what-is-animal-research#Animal%20research%20v%20animal%20testing. Accessed 11 Aug. 2026.

[2]"Problems with Animal Research." American Anti-Vivisection Society, 16 Apr. 2025, aavs.org/animals-science/problems-animal-research/.

[3]"Animal Testing: Animals Used in Experiments." *PETA*, 22 Aug. 2025, www.peta.org/issues/animals-used-for-experimentation/.

[4]"Animal Welfare Act." Animal Welfare Act | National Agricultural Library, www.nal.usda.gov/animal-health-and-welfare/animal-welfare-act. Accessed 11 Aug. 2026.
[5]"PHS Policy." National Institutes of Health, U.S. Department of Health and Human Services, grants.nih.gov/policy-and-compliance/policy-topics/animal-welfare/laws-regulations/phs-policy. Accessed 11 Aug. 2026.
[6]Franco, Nuno Henrique. "Animal Experiments in Biomedical Research: A Historical Perspective." MDPI, MDPI, 19 Mar. 2013, www.mdpi.com/2076-2615/3/1/238.
[7]Nutton, Vivian. Galen | Biography, Achievements, & Facts | Britannica, 26 June 2026, www.britannica.com/biography/Galen.
[8]Thomas, Evan. "Descartes on the Animal Within, and the Animals Without." Canadian Journal of Philosophy 50.8 (2020): 999–1014. Web.
[9]Locke, John. Some thoughts concerning education.. Sherwood, Neely, and Jones, 1809.
[10]Bates, A. W. H. "Vivisection, Virtue, and the Law in the Nineteenth Century." Www.Ncbi.Nlm.Nih.Gov, Palgrave Macmillan, 2017, https://www.ncbi.nlm.nih.gov/books/NBK513716/. Accessed 12 Aug. 2026.
[11]Finn, Michael A., and James F. Stark. "Medical Science and the Cruelty to Animals Act 1876: A Re-Examination of Anti-Vivisectionism in Provincial Britain." Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, vol. 49, Feb. 2015, pp. 12–23, https://doi.org/10.1016/j.shpsc.2014.10.007.
[12]Jorie Braunold. "The American Medical Association on the Ethics of Vivisection, 1880-1950." AMA Journal of Ethics, vol. 26, no. 9, American Medical Association, Sept. 2024, pp. 730–36, https://doi.org/10.1001/amajethics.2024.730. Accessed 12 Aug. 2026.
[13]Strauss, Mark. "Americans Are Divided over the Use of Animals in Scientific Research." Pew Research Center, 16 Aug. 2018, https://www.pewresearch.org/short-reads/2018/08/16/americans-are-divided-over-the-use-of-animals-in-scientific-research/. Accessed 12 Aug. 2026.
[14]Ipsos. "Most Americans Support Vet-Supervised Animal Research." Ipsos, 13 July 2026, https://www.ipsos.com/en-us/most-americans-support-vet-supervised-animal-research. Accessed 12 Aug. 2026.
[15]The Ethics of Research Involving Animals, Nuffield Council on Bioethics, 2005, pp. 239–58, https://cdn.nuffieldbioethics.org/wp-content/uploads/The-ethics-of-research-involving-animals-full-report-2.pdf. Accessed 12 Aug. 2026.
[16]Johns Hopkins University. "Researchers Map Genome for Cats, Dolphins, Birds, and Dozens of Other Animals." The Hub, HUB, 29 Jan. 2024, https://hub.jhu.edu/2024/01/29/researchers-map-51-vertebrate-genomes/. Accessed 12 Aug. 2026.
[17]"Animal Models." Harvardmed, 2025, https://hms.harvard.edu/research/animal-research/animal-models. Accessed 13 Aug. 2026.
[18]National Association for Biomedical Research. "The Importance of Animal Testing in Biomedical Research." Www.Nabr.Org, National Association for Biomedical Research, 2020, https://www.nabr.org/biomedical-research/importance-biomedical-research. Accessed 12 Aug. 2026.
[19]National Human Genome Research Institute. "Why Mouse Matters." Genome.Gov, 23 July 2010, https://www.genome.gov/10001345/importance-of-mouse-genome. Accessed 12 Aug. 2026.
[20]Bryda, Elizabeth C. "The Mighty Mouse: the impact of rodents on advances in biomedical

research." Missouri medicine vol. 110,3 (2013): 207-11.
[21]HUERKAMP.MICHAEL. "Animals Used in Research." Veterinary Manual, MSD Veterinary Manual, 2019, https://www.merckvetmanual.com/exotic-and-laboratory-animals/laboratory-animals/animals-used-in-research. Accessed 13 Aug. 2026.
[22]Hickman, D. L., et al. "Commonly Used Animal Models." Principles of Animal Research for Graduate and Undergraduate Students, vol. 7, no. PMC7150119, Nov. 2016, pp. 117–75, https://doi.org/10.1016/B978-0-12-802151-4.00007-4. Accessed 13 Aug. 2026.
[23]Understanding Animal Research. "Rabbit." Understanding Animal Research, 2021, https://www.understandinganimalresearch.org.uk/using-animals-in-scientific-research/animal-research-species/rabbit. Accessed 13 Aug. 2026.
[24]Esteves, Pedro J., et al. "The Wide Utility of Rabbits as Models of Human Diseases." Experimental & Molecular Medicine, vol. 50, no. 5, Springer Science and Business Media LLC, May 2018, pp. 1–10, https://doi.org/10.1038/s12276-018-0094-1. Accessed 13 Aug. 2026.
[25]Qi, Zhizhen, et al. "Comparison of Mouse, Guinea Pig and Rabbit Models for Evaluation of Plague Subunit Vaccine F1 + rV270." Vaccine, vol. 28, no. 6, Elsevier, Mar. 2009, pp. 1655–60, https://doi.org/10.1016/j.vaccine.2009.02.078. Accessed 13 Aug. 2026.
[26]Canizo, Jesica Romina, et al. "The Guinea Pig Serves as an Alternative Model to Study Human Preimplantation Development." Nature Cell Biology, vol. 27, no. 4, Springer Science and Business Media LLC, Apr. 2025, pp. 696–710, https://doi.org/10.1038/s41556-025-01642-9. Accessed 13 Aug. 2026.
[27]Morrison, Janna L., et al. "Guinea Pig Models for Translation of the Developmental Origins of Health and Disease Hypothesis into the Clinic." The Journal of Physiology, vol. 596, no. 23, Wiley, May 2018, pp. 5535–69, https://doi.org/10.1113/jp274948. Accessed 13 Aug. 2026.
[28]"Housing and Husbandry: Rabbit | Nc3rs." Nc3rs.Org.Uk, 26 Mar. 2014, https://nc3rs.org.uk/3rs-resources/housing-and-husbandry-rabbit. Accessed 16 Aug. 2026.
[29]Soares, João, et al. "The Rabbit as an Animal Model to Study Innate Immunity Genes: Is It Better than Mice?" Frontiers in Immunology, vol. 13, Frontiers Media SA, Sept. 2022, https://doi.org/10.3389/fimmu.2022.981815. Accessed 16 Aug. 2026.
[30]Lee, Erin S, et al. "Cationic Liposome-Oligonucleotide Complex as an Alternative Adjuvant for Polyclonal Antibody Production in New Zealand White Rabbits (Oryctolaguscuniculus)." Comparative Medicine, vol. 67, no. 6, Jan. 2017, pp. 498–503, https://pubmed.ncbi.nlm.nih.gov/29212581/. Accessed 16 Aug. 2026.
[31]Howe, Kerstin, et al. "The Zebrafish Reference Genome Sequence and Its Relationship to the Human Genome." Nature, vol. 496, no. 7446, Apr. 2013, pp. 498–503, https://doi.org/10.1038/nature12111. Accessed 16 Aug. 2026.
[32]Burggren, W. W., and S. Warburton. "Amphibians as Animal Models for Laboratory Research in Physiology." ILAR Journal, vol. 48, no. 3, Jan. 2007, pp. 260–69, https://doi.org/10.1093/ilar.48.3.260. Accessed 16 Aug. 2026.
[33]Hopkins, W. A. "Amphibians as Models for Studying Environmental Change." ILAR Journal, vol. 48, no. 3, Jan. 2007, pp. 270–77, https://doi.org/10.1093/ilar.48.3.270. Accessed 16 Aug. 2026.
[34]Seebacher, Frank, and Craig E. Franklin. "Physiological Mechanisms of Thermoregulation in Reptiles: A Review." Journal of Comparative Physiology B, vol. 175, no. 8, July 2005, pp. 533–41, https://doi.org/10.1007/s00360-005-0007-1.

[35]"Genome Sequence Marks Big Leap Forward for Frog Researchers." URMC Newsroom, 28 Apr. 2010, https://www.urmc.rochester.edu/news/story/genome-sequence-marks-big-leap-forward-for-frog-researchers. Accessed 16 Aug. 2026.
[36]Dowling, Herndon G, and George R Zug. "Reptile - Circulatory System." Encyclopædia Britannica, 2018, https://www.britannica.com/animal/reptile/Circulatory-system. Accessed 16 Aug. 2026.
[37]Jensen, Bjarke, and Vincent M Christoffels. "Reptiles as a Model System to Study Heart Development." Cold Spring Harbor perspectives in biology vol. 12,5 a037226. 1 May. 2020, doi:10.1101/cshperspect.a037226
[38]Kaiser, Pete. "The Long View: A Bright Past, a Brighter Future? Forty Years of Chicken Immunology Pre- and Post-Genome." Avian Pathology, vol. 41, no. 6, Informa UK Limited, Dec. 2012, pp. 511–18, https://doi.org/10.1080/03079457.2012.735359. Accessed 16 Aug. 2026.
[39]Kaufman, Jim. "The Simple Chicken Major Histocompatibility Complex: Life and Death in the Face of Pathogens and Vaccines." Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, edited by C. R. M. Bangham et al., vol. 355, no. 1400, The Royal Society, Aug. 2000, pp. 1077–84, https://doi.org/10.1098/rstb.2000.0645. Accessed 16 Aug. 2026.
[40]Kmecick, Melyssa, et al. "Morphological Evidence of Neurotoxic Effects in Chicken Embryos after Exposure to Perfluorooctanoic Acid (PFOA) and Inorganic Cadmium." Toxicology, vol. 427, Elsevier, Sept. 2019, p. 152286, https://doi.org/10.1016/j.tox.2019.152286. Accessed 17 Aug. 2026.
[41]Scanes, Colin & Mcnabb, Anne. (2003). Avian Models for Research in Toxicology and Endocrine Disruption. Avian and Poultry Biology Reviews. 14. 21-52. 10.3184/147020603783727021.
[42]CDC. "How Influenza (Flu) Vaccines Are Made." Influenza (Flu), 30 Sept. 2024, https://www.cdc.gov/flu/vaccine-process/. Accessed 17 Aug. 2026.
[43]"Researchers Compare Chicken, Human Genomes." Genome.Gov, 2011, https://www.genome.gov/12514316/2004-release-researchers-compare-chicken-human-genomes. Accessed 17 Aug. 2026.
[44]Small, D L, and A M Buchan. "Animal models." British medical bulletin vol. 56,2 (2000): 307-17. doi:10.1258/0007142001903238
[45]"Swine Models 2024: Centers and Research Resources." Office of Research Infrastructure Programs, National Institutes of Health, 2024, orip.nih.gov/sites/default/files/Swine_Models_2024.pdf. Accessed 17 Aug. 2026.
[46]Xi, Jiahui, et al. "Genetically Engineered Pigs for Xenotransplantation: Hopes and Challenges." Frontiers in Cell and Developmental Biology, vol. 10, no. 1093534, Jan. 2023, https://doi.org/10.3389/fcell.2022.1093534. Accessed 17 Aug. 2026.
[47]Alex, Aneesh, et al. "In Vivo Characterization of Minipig Skin as a Model for Dermatological Research Using Multiphoton Microscopy." Experimental Dermatology, vol. 29, no. 10, Wiley-Blackwell, Aug. 2020, pp. 953–60, https://doi.org/10.1111/exd.14152.
[48]Stanford Medicine. "Facts and Myths." Animal Research at Stanford, 2019, https://med.stanford.edu/animalresearch/facts-and-myths.html. Accessed 17 Aug. 2026.
[49]Zheng, Chengli et al. "Primary Evaluation of Three-Dimensional Printing-Guided Endodontics in the Dog Maxillary." Veterinary sciences vol. 12,7 665. 14 Jul. 2025, doi:10.3390/vetsci12070665. Accessed 17 Aug. 2026.
[50]Griffin, Brenda, and Henry J. Baker. "Domestic Cats as Laboratory Animals."

Laboratory Animal Medicine, 2002, pp. 459–82, https://doi.org/10.1016/B978-012263951-7/50015-6. Accessed 17 Aug. 2026.
[51]American Museum of Natural History. "DNA: Comparing Humans and Chimps." American Museum of Natural History, American Museum of Natural History, 2019, https://www.amnh.org/exhibitions/permanent/human-origins/understanding-our-past/dna-comparing-humans-and-chimps. Accessed 12 Aug. 2026.
[52]"Medical Testing on Chimpanzees No Longer Allowed, the National Institutes of Health Says." PBS News, 13 Dec. 2015, https://www.pbs.org/newshour/show/medical-testing-on-chimpanzees-no-longer-allowed-the-national-institutes-of-health-says. Accessed 17 Aug. 2026.
[53]"Nonhuman Primates | Animal Welfare Institute." Animal Welfare Institute, 15 Apr. 2026, https://awionline.org/programs/animals-in-laboratories/strengthening-laws-regulations-and-enforcement/nonhuman-primates/. Accessed 17 Aug. 2026.
[54]Haigwood, Nancy. "The Critical Role of Nonhuman Primates in Medical Research." OHSU News, https://news.ohsu.edu/2017/03/13/the-critical-role-of-nonhuman-primates-in-medical-research. Accessed 17 Aug. 2026.
[55]"Animal Care and Facilities." Animal Research at Stanford, 2019, https://med.stanford.edu/animalresearch/animal-care-and-facilities.html. Accessed 17 Aug. 2026.
[56]"Regulatory Oversight." Animal Research at Stanford, https://med.stanford.edu/animalresearch/regulatory-oversight.html. Accessed 17 Aug. 2026.
[57]Harvard.Edu, 2024, https://hms.harvard.edu/research/animal-research/animal-welfare. Accessed 17 Aug. 2026.
[58]"Animal Welfare Act." National Agricultural Library, U.S. Department of Agriculture, 2023, https://www.nal.usda.gov/animal-health-and-welfare/animal-welfare-act. Accessed 17 Aug. 2026.
[59]"PHS Policy | Grants & Funding." Nih.Gov, 2019, https://grants.nih.gov/policy-and-compliance/policy-topics/animal-welfare/laws-regulations/phs-policy. Accessed 17 Aug. 2026.
[60]Kiani, Aysha Karim. "Ethical Considerations Regarding Animal Experimentation." Journal of Preventive Medicine and Hygiene, vol. 63, no. 2 Suppl 3, Oct. 2022, pp. E255–66, https://doi.org/10.15167/2421-4248/jpmh2022.63.2S3.2768.
[61]Ferdowsian, Hope R., and Nancy Beck. "Ethical and Scientific Considerations Regarding Animal Testing and Research." PLoS ONE, edited by Catriona J. MacCallum, vol. 6, no. 9, Sept. 2011, p. e24059, https://doi.org/10.1371/journal.pone.0024059. Accessed 17 Aug. 2026.
[62]Lee, Kook Hyun, et al. "The 'R' Principles in Laboratory Animal Experiments." Laboratory Animal Research, vol. 36, no. 1, Dec. 2020, https://doi.org/10.1186/s42826-020-00078-6.
[63]Liu, Jimin, et al. "Practical Implementation and Impact of the 4R Principles in Ethnopharmacology: Pursuing a More Humane Approach to Research." Frontiers in Pharmacology, vol. 16, Frontiers Media, Mar. 2025, https://doi.org/10.3389/fphar.2025.1543316.
[64]Kang, Meili, et al. "A Review of the Ethical Use of Animals in Functional Experimental Research in China Based on the 'Four R' Principles of Reduction, Replacement, Refinement, and Responsibility." Medical Science Monitor, vol. 29, Dec. 2022, https://doi.org/10.12659/msm.938807.
[65]Russell, W, and R Burch. The Principles of Humane Experimental Technique. https://caat.publichealth.jhu.edu/wp-content/uploads/2024/08/The-Control-of-the-Proximate-especially-Behavioral-Environment.pdf. Accessed 18 Aug. 2026.

[66]Petkov, Christopher I., et al. "Unified Ethical Principles and an Animal Research 'Helsinki' Declaration as Foundations for International Collaboration." Current Research in Neurobiology, vol. 3, Elsevier BV, 2022, p. 100060, https://doi.org/10.1016/j.crneur.2022.100060. Accessed 18 Aug. 2026.
[67]"Principles of Animal Research Ethics." Google Books, 2026, https://books.google.com/books?hl=en&lr=&id=FPDGDwAAQBAJ&oi=fnd&pg=PP1&ots=GFtsDfZNBL&sig=z4je_xCyO7q6hJdfQZXA_cUZyTE#v=onepage&q&f=false. Accessed 18 Aug. 2026.
[68]"Institutional Animal Care and Use Committee (IACUC) | Icahn School of Medicine." Icahn School of Medicine at Mount Sinai, 2025, https://icahn.mssm.edu/research/iacuc. Accessed 18 Aug. 2026.
[69]"How Animal Research Is Regulated in the U.S.." Default, https://www.physiology.org/career/policy-advocacy/animal-research/how-animal-research-is-regulated?SSO=Y. Accessed 18 Aug. 2026.
[70]"Our Resources | NC3Rs." Nc3rs.Org.Uk, https://nc3rs.org.uk/3rs-resources. Accessed 18 Aug. 2026.
[71]"About." ARRIVE Guidelines, 3 June 2020, https://arriveguidelines.org/about. Accessed 18 Aug. 2026.
[72]Jain, Abhishek K., et al. "Models and Methods for In Vitro Toxicity." In Vitro Toxicology, 2018, pp. 45–65, https://doi.org/10.1016/b978-0-12-804667-8.00003-1. Accessed 18 Aug. 2026.
[73]"In Silico - an Overview | ScienceDirect Topics." Www.Sciencedirect.Com, https://www.sciencedirect.com/topics/medicine-and-dentistry/in-silico. Accessed 18 Aug. 2026.
[74]Harvard Medical School. "Alternatives to Animal Research." Harvard.Edu, 2024, https://hms.harvard.edu/research/animal-research/alternatives-animal-research. Accessed 18 Aug. 2026.
[75]"Myths about Animal Research." Harvard.Edu, 2024, https://hms.harvard.edu/research/animal-research/myths-about-animal-research. Accessed 18 Aug. 2026.
[76]American Anti-Vivisection Society. "Problems with Animal Research." American Anti-Vivisection Society, American Anti-Vivisection Society, 2025, https://aavs.org/animals-science/problems-animal-research/. Accessed 18 Aug. 2026.
[77]PETA. "Animals Used for Experimentation | PETA." PETA, 2024, https://www.peta.org/issues/animals-used-for-experimentation/ . Accessed 18 Aug. 2026.
[78]Medical School, Harvard. "What Animal Research Has Given Us." Harvard.Edu, 2024, https://hms.harvard.edu/research/animal-research/what-animal-research-has-given-us. Accessed 18 Aug. 2026.
[79]"Stanford Discoveries." Animal Research at Stanford, 30 Oct. 2018, https://med.stanford.edu/animalresearch/stanford-discoveries.html. Accessed 18 Aug. 2026.
[80]Anderson, Jamie E., et al. "Fetal Repair of Surgically-Created Gastroschisis in a Lamb Model May Improve Postnatal Feeding Tolerance and Intestinal Motility: A Pilot Study." Journal of Pediatric Surgery, vol. 54, no. 10, W.B. Saunders, Jan. 2019, pp. 2061–68, https://doi.org/10.1016/j.jpedsurg.2018.11.021. Accessed 18 Aug. 2026.
[81]Brownlee, Rachel D, et al. "Blood Pressure Reference Intervals for Ketamine-Sedated Rhesus Macaques (Macaca Mulatta)." Journal of the American Association for Laboratory Animal Science, 2020, https://doi.org/10.30802/aalas-jaalas-19-000072.
[82]UCI Office of Research. "Euthanasia of Research Animals." UCI Office of Research, 2024, https://research.uci.edu/animal-care-and-use/poli

cies-and-guidance/euthanasia-of-research-animals/. Accessed 19 Aug. 2026.
[83]"AVMA Guidelines for the Euthanasia of Animals." American Veterinary Medical Association, 2020, https://www.avma.org/resources-tools/avma-policies/avma-guidelines-euthanasia-animals. Accessed 19 Aug. 2026.
[84]"Experimental Infection of Mice with Veronaea Botryosa as a Model for Human Phaeohyphomycosis." Google.Com, 2019, https://scholar.google.com/citations?view_op=view_citation&hl=en&user=zUadQrkAAAAJ&citation_for_view=zUadQrkAAAAJ:d1gkVwhDpl0C. Accessed 19 Aug. 2026.
[85]"Humane Killing | NC3Rs." Nc3rs.Org.Uk, Humane killing | NC3Rs, 21 Oct. 2013, https://nc3rs.org.uk/3rs-resource-library/humane-killing. Accessed 19 Aug. 2026.
[86]Stop Euthanizing Animals After Experiments. "Stop Euthanizing Animals After Experiments." Congressman Brian Mast, 29 Mar. 2023, https://mast.house.gov/2023/3/stop-euthanizing-animals-after-experiments. Accessed 19 Aug. 2026.
[87]Rafaiee, Raheleh, et al. "Ethical Considerations and Challenges Associated with Euthanasia in Laboratory Animal Research." Journal of Medical Ethics and History of Medicine, Knowledge E DMCC, May 2025, https://doi.org/10.18502/jmehm.v17i11.18663. Accessed 19 Aug. 2026.
[88]"Adoption of Laboratory Animals after Research | Grants & Funding." Nih.Gov, 2019, https://grants.nih.gov/policy-and-compliance/policy-topics/air/adoption. Accessed 19 Aug. 2026.
[89]Juan. "Animals Used in Research - MSPCA-Angell." MSPCA-Angell, 26 Mar. 2026, https://www.mspca.org/advocacy-issues/animals-used-in-research/. Accessed 19 Aug. 2026.
[90]Elliott, James J., et al. "Management of Research Animal Breeding Colonies." PubMed, edited by Robert H. Weichbrod et al., 2nd ed., CRC Press/Taylor & Francis, 2018, https://www.ncbi.nlm.nih.gov/books/NBK500423/. Accessed 19 Aug. 2026.
[91]"Laboratory Animal Resource Center (LARC)." Ucsf.Edu, 2026, https://larc.ucsf.edu/acquisition-procedure. Accessed 19 Aug. 2026.
[92]Vanhoose, Hannah. Comparative Analysis: Animal Research versus Animal Testing Comparative Analysis: Animal Research versus Animal Testing. https://digitalcommons.longwood.edu/cgi/viewcontent.cgi?article=1049&context=rci_fall. Accessed 19 Aug. 2026.
[93]National Research Council (US) and Institute of Medicine (US) Committee on the Use of Laboratory Animals in Biomedical and Behavioral Research. "Patterns of Animal Use." Nih.Gov, National Academies Press (US), 2014, https://www.ncbi.nlm.nih.gov/books/NBK218261/. Accessed 19 Aug. 2026.
[94]"IRAC Recommendation on LD50 Testing: Interagency Research Animal Committee (IRAC)." ILAR Journal, vol. 35, no. 3–4, Oxford University Press (OUP), Jan. 1993, pp. 56–58, https://doi.org/10.1093/ilar.35.3-4.56. Accessed 19 Aug. 2026.
[95]Wilson, Samantha L., et al. "An Overview of Current Techniques for Ocular Toxicity Testing." Toxicology, vol. 327, Jan. 2015, pp. 32–46, https://doi.org/10.1016/j.tox.2014.11.003.
[96]"Repeat Dose Toxicity Study - an Overview | ScienceDirect Topics." Www.Sciencedirect.Com, https://www.sciencedirect.com/topics/pharmacology-toxicology-and-pharmaceutical-science/repeat-dose-toxicity-study. Accessed 19 Aug. 2026.
[97]Erhirhie, Earnest Oghenesuvwe, et al. "Advances in Acute Toxicity Testing: Strengths, Weaknesses and Regulatory Acceptance." Interdisciplinary Toxicology, vol. 11, no. 1, Walter de Gruyter GmbH, May 2018, pp. 5–12,

https://doi.org/10.2478/intox-2018-0001. Accessed 19 Aug. 2026.
[98]Humane World For Animals. "Cosmetics Animal Testing FAQ." Humane World, 7 Feb. 2025, https://www.humaneworld.org/en/issue/cosmetics-animal-testing-FAQ. Accessed 19 Aug. 2026.
[99]"What Animal Research Has Given Us." Harvardmed, 2025, https://hms.harvard.edu/research/animal-research/what-animal-research-has-given-us. Accessed 19 Aug. 2026.
[100]"Therapies & Treatments – Cancer." Fbresearch.Org, 2016, https://fbresearch.org/therapies-treatments/cancer. Accessed 19 Aug. 2026.
[101]"Novel Cancer Vaccine Offers New Hope for Dogs — and Those Who Love Them." Yale News, 5 Mar. 2024, https://news.yale.edu/2024/03/05/novel-cancer-vaccine-offers-new-hope-dogs-and-those-who-love-them. Accessed 19 Aug. 2026.
[102]Thamm, Douglas H. "Canine Cancer: Strategies in Experimental Therapeutics." Frontiers in Oncology, vol. 9, Frontiers Media SA, Nov. 2019, https://doi.org/10.3389/fonc.2019.01257. Accessed 19 Aug. 2026.
[103]Gardner, Heather L., et al. "Dogs as a Model for Cancer." Annual Review of Animal Biosciences, vol. 4, no. 1, Annual Reviews, Feb. 2016, pp. 199–222, https://doi.org/10.1146/annurev-animal-022114-110911. Accessed 19 Aug. 2026.
[104]"Polio." Understanding Animal Research, 2026, https://www.understandinganimalresearch.org.uk/human-diseases/polio-vaccine. Accessed 19 Aug. 2026.
[105]CDC. "About Polio in the United States." Polio, 25 Sept. 2024, https://www.cdc.gov/polio/about/index.html. Accessed 19 Aug. 2026.
[106]National Library of Medicine. "Benefits Derived from the Use of Animals." Nih.Gov, National Academies Press (US), 2019, https://www.ncbi.nlm.nih.gov/books/NBK218274/. Accessed 19 Aug. 2026.
[107]Yamamoto, Janet, et al. "Feline Immunodeficiency Virus Model for Designing HIV/AIDS Vaccines." Current HIV Research, vol. 8, no. 1, Jan. 2010, pp. 14–25, https://doi.org/10.2174/157016210790416361.
[108]Moretti, Sonia, et al. "Advances in SIV/SHIV Non-Human Primate Models of NeuroAIDS." Pathogens, vol. 10, no. 8, Multidisciplinary Digital Publishing Institute, Aug. 2021, pp. 1018–18, https://doi.org/10.3390/pathogens10081018.
[109]Kumar, Nitasha, et al. "Animal Models to Achieve an HIV Cure." Current Opinion in HIV and AIDS, vol. 11, no. 4, Ovid Technologies (Wolters Kluwer Health), July 2016, pp. 432–41, https://doi.org/10.1097/coh.0000000000000290. Accessed 19 Aug. 2026.
[110]"A Surprising Triple Treatment Completely Clears Hiv in Newborn Primates." ScienceDaily, 2026, https://www.sciencedaily.com/releases/2026/08/260812015200.htm. Accessed 19 Aug. 2026.
[111]"A Novel Gene-Therapy Approach to 'Functionally Cure' HIV Succeeds in Some Monkeys." AAAS Articles DO Group, American Association for the Advancement of Science (AAAS), Apr. 2026, https://doi.org/10.1126/science.zorh0qr. Accessed 19 Aug. 2026.
[112]"Therapies & Treatments – Neurodegenerative Diseases." Fbresearch.Org, 2016, https://fbresearch.org/therapies-treatments/neurodegenerative-diseases. Accessed 19 Aug. 2026.
[113]Goldman, Bruce. "Alzheimer's-Implicated Protein Tau Scrambles Nerve Cells' Powerpacks." News Center, 6 Aug. 2026, https://med.stanford.edu/news/all-news/2026/08/tau-alzheimers.html. Accessed 19 Aug. 2026.

[114]Foundation for Biomedical Research. "Animal Research Achievements." Foundation for Biomedical Research, 2016, https://fbresearch.org/medical-advances/animal-research-achievements. Accessed 19 Aug. 2026.
[115]Wang, Shen, et al. "COVID-19 Animal Models and Vaccines: Current Landscape and Future Prospects." Vaccines, vol. 9, no. 10, Sept. 2021, p. 1082, https://doi.org/10.3390/vaccines9101082.
[116]Huffman, Carl. "Alcmaeon." Stanford.Edu, 10 Apr. 2003, https://plato.stanford.edu/archives/spr2026/entries/alcmaeon/#Pyth.
[117]Alstrup, Aage Kristian Olsen. "Animal Research in the Age of Antiquity - Danish 3R-Center." En.3rcenter.Dk, https://en.3rcenter.dk/laboratory-animals/history-of-animal-testing/animal-research-in-the-age-of-antiquity. Accessed 20 Sept. 2026.
[118]"Diocles of Carystus | EBSCO." EBSCO Information Services, Inc. | Www.Ebsco.Com, 2023, https://www.ebsco.com/research-starters/biography/diocles-carystus.
[119]Franco, Nuno. "Animal Experiments in Biomedical Research: A Historical Perspective." Animals, vol. 3, no. 1, Mar. 2013, pp. 238–73, https://doi.org/10.3390/ani3010238.
[120]Browning, Heather, and Jonathan Birch. "Teaching & Learning Guide for: Animal Sentience." Philosophy Compass, vol. 17, no. 11, Oct. 2022, https://doi.org/10.1111/phc3.12878.
[121]National Library of Medicine. "Homology: Orthologs and Paralogs." Www.Nlm.Nih.Gov, https://www.nlm.nih.gov/ncbi/workshops/2023-08_BLAST_evol/ortho_para.html. Accessed 20 Sept. 2026.
[122]National Human Genome Research Institute. "Haplotype." Genome.Gov, 2019, https://www.genome.gov/genetics-glossary/haplotype.
[123]Tregaskes, Clive A., and Jim Kaufman. "Chickens as a Simple System for Scientific Discovery: The Example of the MHC." Molecular Immunology, vol. 135, July 2021, pp. 12–20, https://doi.org/10.1016/j.molimm.2021.03.019.
[124]Shimizu, A, and G Kanzaki. "Xenotransplantation - an Overview | ScienceDirect Topics." Www.Sciencedirect.Com, 2014, https://www.sciencedirect.com/topics/medicine-and-dentistry/xenotransplantation.
[125]"Definition of LENTIVIRUS." Www.Merriam-Webster.Com, https://www.merriam-webster.com/dictionary/lentivirus. Accessed 20 Sept. 2026.
[126]"Animal Welfare Act Quick Reference Guides | National Agricultural Library." Www.Nal.Usda.Gov, https://www.nal.usda.gov/animal-health-and-welfare/animal-welfare-act-quick-reference-guides. Accessed 22 Sept. 2026.
[127]Public Health Service Policy on Humane Care and Use of Laboratory Animals. https://grants.nih.gov/sites/default/files/PHSPolicyLabAnimals.pdf. Accessed 22 Sept. 2026.
[128]AALAS Career Center. "Animal Welfare Act and PHS Policy Compliance — a Practical Guide for Lab Animal Science Professionals." American Association for Laboratory Animal Science, AALAS Career Center, 22 June 2026, https://careercenter.aalas.org/career-resources/on-the-job-3/animal-welfare-act-and-phs-policy-compliance-a-practical-guide-for-lab-animal-science-professionals-116.
[129]"Higher Animals Noun - Definition, Pictures, Pronunciation and Usage Notes | Oxford Advanced Learner's Dictionary at OxfordLearnersDictionaries.Com." Oxfordlearnersdictionaries.Com, 2026, https://www.oxfordlearnersdictionaries.com/definition/english/higher-animals.
[130]Arguments Against Animal Testing. https://www.crueltyfreeinternational.org/wp-cont

ent/uploads/2025/11/Arguments-Against-Animal-Testing.pdf. Accessed 22 Sept. 2026.

[131]"Animal Research | Research, Innovation & Impact." Research.Missouri.Edu, https://research.missouri.edu/animal-research. Accessed 22 Sept. 2026.

[132]"Fully Accredited GLP Lab | Noble Life Sciences Facilities." Noble Life Sciences, 8 Nov. 2022, https://noblelifesci.com/about-noble/facilities/.

[133]UCI Office of Research. "Euthanasia of Research Animals." UCI Office of Research, 2024, https://research.uci.edu/animal-care-and-use/policies-and-guidance/euthanasia-of-research-animals/.

[134]Reilly, J. *Euthanasia of Animals Used for Scientific Purposes Edited By.* 2001, https://nre.tas.gov.au/Documents/ANZCCART-Guidelines-for-Euthanasia-of-Animals.pdf. Accessed 22 Sept. 2026.

[135]Sivula, Christine P., and Mark A. Suckow. Euthanasia." *PubMed,* edited by Robert H. Weichbrod et al., 2nd ed., CRC Press/Taylor & Francis, 2018, https://www.ncbi.nlm.nih.gov/books/NBK500441/. Accessed 22 Sept. 2026.

Journal of Ethics in Scientific and Technological Innovations

Assessing the Ethics of Trials That Depend on Risk: Studying People Who Inject Drugs

Luke Zhang[1]
[1]*Southfield Christian High School, Southfield, MI 48034*

Abstract

Certain clinical trials rely on participants performing a dangerous activity. Recently, the clearest cases are HIV prevention trials in people who inject drugs. Studies need new infections to measure variables and new infections only come from continued injection of drugs. Between the years of 2004 and 2005, activists forced three of these trials to be halted or suspended, arguing that researchers were promoting risk for participants. This review examines whether this objection is valid. I will argue it does not, but that the standard defense of these trials fails is exact enough. Excluding people who inject drugs does not protect them, but simply removes them from the evidence while they keep acquiring HIV. The core problem is narrower- researchers leave out prevention tools that already work, then treat the infections that follow as if they came from nowhere. The effects of tenofovir were unknown, so testing it was fair, though unknown effectiveness alone does not make a trial entirely ethical. In Bangkok specifically, the gap was withheld sterile injecting equipment, not opioid substitution therapy, which participants did receive; more broadly, however, omitting opioid substitution therapy, which cuts HIV acquisition by half, is not acceptable. I propose three conditions for continuation and argue that the strongest designs, exemplified by HPTN 074, put the prevention itself on trial, so the study benefits participants through the same procedure that produces its result.

1. Introduction

In July of 2004, the prime minister of Cambodia stopped a trial of tenofovir for HIV prevention in Phnom Penh sex workers. Later, in February of 2005, Cameroon's health minister suspended a similar trial. In March 2005, Thai drug user organizations protested a third trial in Bangkok.[1,2] The objections raised across all three studies were consistent: participants were recruited because they were high risk individuals but the trials did not give any meaningful ways to reduce that risk. More concretely, activists in Cambodia and Cameroon cited inadequate counseling and a protocol that limited those who could receive antiretroviral therapy to only those who officially met WHO criteria for AIDS.[1] Additionally, activists in Bangkok alleged that the trial refused to supply sterile injecting equipment while running a placebo arm.[2]

People who inject drugs are a population for whom standard HIV prevention advice, such as abstaining from injection or relying solely on condom use, is often neither feasible nor sufficient, so trials in this group study interventions—pre-exposure prophylaxis, opioid substitution therapy, sterile syringe access—that target injection-related transmission directly. This population is also frequently criminalized, stigmatized, and excluded from routine healthcare, which makes its members simultaneously among those most in need of effective prevention research and among those most vulnerable to being exploited by it.

However, the Bangkok trial proceeded. It enrolled 2,413 participants and reported in 2013 that daily oral tenofovir reduced HIV incidence by 48.9% (95% CI 9.6–72.2).[3] This study has laid the foundation for the support of pre-exposure prophylaxis to all who inject drugs. Had this trial been stopped, the recommendation would not exist and the population that it now protects would be at risk.

Herein lies the tension: this was a trial whose scientific value depended, if not solely, heavily, on participants continuing a risky behavior. While looking exploitative, halting it does not end the behavior; it only ends the knowledge. Haire treats the trials as a failure of

researcher obligation[12]while Singh and Mills treat them as a failure of community engagement that cost the field years of evidence.[1] Neither position is complete nor incorrect.

Below, I set out the case against these trials then, the case for continuing them, and a verdict: that the objection identifies a real defect but prescribes the wrong remedy.

2. The Case Against

2.1 Exploitation and Coercion

The strongest version of the objection does not claim that participants were forced or deceived. Wertheimer's account is unique in the perspective it offers - a translation can be voluntary, informed, and beneficial to both parties, yet still be exploitative, because exploitation is a claim about the fairness of the terms, not consent.[19] Therefore, a project can be procedurally sound yet take more from participants than it gives back.

This characterization of exploitation is the primary charge against the 2004 trials because participants possessed the one thing the study could not manufacture: ongoing exposure to HIV. However, they only received in exchange, a drug of unknown value, and a coin flip's chance as to whether they got it, and an arbitrary level of prevention support the sponsors chose. Additionally, Haire argues that when researchers watch participants seroconvert, the obligation to provide treatment and other health benefits follows from reciprocal justice rather than from true sponsor policy.[12] In the 2004 trials and similar experiments, the terms were set by the party with all the power, making them exploitative in nature, despite the consent of the participants.

2.2 Issues with the Standard of Prevention

Philpott and colleagues document that the field entered these trials without an agreed standard for what prevention participants are owed, and that "state of the art" was interpreted differently by sponsors, host countries, and advocates.[13] Future work by Dawson and Zwerski showed that the disparity continued and trial designers were making calls case by case.[14] The undefined nature of the obligation made it easy to define it downward, applying the minimum possible standard of care in the worst situations.

2.3 Consent Under Criminalization

The third component concerns the conditions under which participants agreed. The Bangkok trial's larger context included a Thai anti-drug campaign associated with roughly 2,800 extrajudicial killings, and participants were recruited at methadone clinics where some feared consequences for refusing.[2] Similar contexts have been observed around the globe. A systematic review of 106 studies found that criminalization of drug use was associated with worse HIV prevention and treatment outcomes in 80% of them.[10] Dawson and colleagues, writing specifically about HIV prevention research in people who inject drugs, are concerned about legal risk as a distinct ethical problem with implications on results.[15]

However, this is not to say consent is impossible under criminalization. It changes what disclosure requires in the way that participants will know what happens to their data if it is subpoenaed, whether enrollment is visible to police or to the clinic controlling their methadone, and that refusal carries no consequence for care they already receive. The separation of powers between the trial administrators and the legal system is required in such contexts as participants will not infer the separation on their own.

.

3. The Case For

3.1 The Necessity of the Research

The core promoting research into high-risk populations relies on the fact that removing high-risk populations from research does not remove them from risk while simultaneously removing them from the data. The decisions that clinicians and public health programs will inevitably have to make about these individuals will be made through extrapolation and guesswork, rather than empirical evidence. For this subset of the population, extrapolation often serves as a poor mechanism due to differences in drug interactions and adherence patterns, as well as competing causes of death.

The 2024 revision of the Declaration of

Helsinki addresses the point directly. Paragraph 19 requires that the harms of excluding groups in situations of vulnerability be weighed against the harms of including them, and paragraph 20 holds that research in such groups is justified when it is "responsive to their health needs and priorities."[4] Emanuel and colleagues make the same point through their requirement of fair subject selection: eligibility criteria should follow scientific necessity, not administrative convenience or discomfort with the population.[5] CIOMS Guideline 15 goes further and warns that treating a group as categorically too vulnerable to enroll denies it the benefits of research.[18] A blanket exclusion would fail all three tests.

The Bangkok trial is the concrete case of the benefits of such trials. It ran, and the recommendation of pre-exposure prophylaxis for people who inject drugs exists because it ran. The Cambodian and Cameroonian trials did not run, and the questions they were designed to answer went unanswered for years.[1]

3.2 The Non-Causation Defense & Limitations

Investigators defend these trials by pointing out that they did not cause the drug use and cannot end it. This is true, and it establishes that the trial's incremental risk is the study procedures rather than the injection itself. I grant the point. It proves less than it is used to prove.

The defense answers the charge that researchers manufacture risk. It does not answer the charge that researchers decline to reduce risk they are positioned to reduce. Richardson and Belsky explain why the second charge sticks. Ancillary care duties arise when participants grant researchers privileged access to their health, and they scale with the depth of that access.[9] A trial that screens for injection frequency and then sees participants monthly for years has that access. It is not a bystander, and the fact that it did not create the danger does not make it one.

4. Verdict

Deciding between these two cases requires a line, and Freedman supplies it. He located the ethical license for randomization in genuine disagreement within the professional community about which arm is better—what he calls clinical equipoise, the state of genuine uncertainty in which no one yet knows which treatment arm is superior.[6] Apply that to Bangkok and the trial splits into two parts that deserve opposite verdicts.

For tenofovir, the disagreement existed. Nobody knew whether daily oral tenofovir prevented HIV in people who inject drugs, which is exactly why the trial was worth running and why a placebo arm was defensible. The 48.9% figure was not knowable in advance, and the wide confidence interval around it (9.6–72.2) shows how close the trial came to answering nothing at all.[3]

For opioid substitution therapy, the disagreement did not exist. A meta-analysis published the year before the trial reported found that opiate substitution treatment was associated with a 54% reduction in HIV acquisition (rate ratio 0.46, 95% CI 0.32–0.67).[7] That is a settled finding being omitted, and no appeal to equipoise covers it.

The Bangkok trial did better on this than its critics allowed, and the detail matters. Participants received methadone, condoms, bleach for cleaning equipment, risk-reduction counseling, and monthly HIV testing. Methadone is opioid substitution therapy, so on the strongest version of the charge the trial complied. What participants did not receive was sterile needles, because Thai narcotics law prohibits distributing them.[3] That is the sharp version of the problem. The gap was not created by the investigators, but the trial ran inside it, and the placebo arm absorbed the consequences.

Here I have to concede something that weakens my own position. The evidence for needle and syringe programmes is not as strong as the evidence for opioid substitution therapy. Aspinall and colleagues found a pooled effect size of 0.66 for the association between needle and syringe programme exposure and HIV transmission, with a confidence interval of 0.43 to 1.01 that crosses one.[16] Restricted to six higher-quality studies the estimate improves to 0.42 (95% CI 0.22–0.81), but the overall picture is weaker than the 54% figure for opioid substitution therapy. A defender of the Bangkok investigators could therefore argue that syringe access sat closer to genuine uncertainty than the

activists claimed, and that the equipoise line does not cleanly condemn the omission.

I believe the argument fails, for two reasons. First, the higher-quality subset is the relevant one, and it points the same way as every mechanistic account of how injection-related HIV transmission works. Second, and more decisively, the investigators did not omit syringes because they judged the evidence insufficient. They omitted them because Thai law forbade distribution.[3] A gap produced by a legal prohibition is not an expression of professional disagreement, and dressing it as one confuses two very different things.

So the verdict. The activists were right that the trials withheld prevention they should have supplied and right that the withholding was hidden behind local standard of care. They were wrong that stopping the trials was the answer, because stopping produced no syringes, no methadone, and no evidence.

5. Three Conditions & Limitations

UNAIDS and WHO set the standard in guidance point 13, which requires that participants receive counseling and access to all state-of-the-art HIV risk reduction methods throughout the trial.[8] "State of the art" is the operative phrase: it is defined by the evidence, not by what the host country currently permits or funds. From that standard and the argument above, three conditions follow.

First, the prevention package meets guidance point 13 in full, including interventions the host country restricts, or the trial does not open at that site. Where local law blocks a proven intervention, the obligation is to negotiate an exemption or supply the intervention through another channel. Failing both, the site is wrong. Second, ancillary care obligations are budgeted and stated in the protocol rather than left to site discretion, which is the practical form of the Richardson and Belsky duty.[9] Third, consent addresses the legal exposure participation creates and separates recruitment from any institution controlling participants' access to treatment.

Three caveats follow just as directly, and I would rather state them than let a reader find them.

The first is a matter of arithmetic. Larney and colleagues surveyed 179 countries with documented injecting drug use and found needle and syringe programmes in 93 of them and opioid substitution therapy in 86, with coverage in most countries falling below WHO benchmarks.[17] Applied strictly, condition one would close roughly half the world to this research, and it would close it precisely where the epidemic is worst and the evidence is thinnest. That is a real cost: a country closed to this research does not stop having an epidemic, and its clinicians are left extrapolating from other populations indefinitely. My answer is that the obligation falls on the trial rather than on the country, so a sponsor can supply what a health system does not. That answer works for underfunding. It does not work for prohibition, and where a state actively forbids the intervention, as Thailand did, I accept the consequence that the trial belongs somewhere else.

The second caveat is that "state of the art" is not self-applying. Philpott and Dawson both show that reasonable people read the phrase differently, and my conditions inherit that indeterminacy.[13,14] The Aspinall numbers above are an example: two careful readers can disagree about whether syringe access had crossed the threshold from open question to settled finding by 2005.

The third caveat is the largest. These are conditions on trials. They do nothing about the criminalization that made the Bangkok context dangerous in the first place, and a trial run perfectly under condition three still enrolls people whose legal exposure the researchers cannot remove.

The stronger move is to change the question. HPTN 074 tested an integrated intervention combining psychosocial counseling with supported referral to antiretroviral therapy and medication for opioid use disorder across Ukraine, Indonesia, and Vietnam. Among 502 participants living with HIV and 806 of their injection partners, the intervention arm roughly doubled antiretroviral therapy uptake, more than halved mortality, and produced zero new infections among injection partners, compared with seven in the standard care arm.[11] The design generates knowledge and benefit from the same procedure. It does not need participants

to keep injecting in order to produce a result, which removes the structural conflict rather than managing it.

This design does not generalize to every case, however. It works because the intervention being tested—integrated care and treatment referral—is itself the benefit, so testing it and delivering it are the same act. A trial like the tenofovir study is different in kind: the question is whether an unproven drug prevents infection, and that question cannot be answered by simply giving everyone the drug, because equipoise, not benefit-delivery, is what makes a placebo arm defensible in the first place. The HPTN 074 model is therefore the answer for questions about care delivery and uptake, where the intervention's value is already established and the open question is how to deliver it. It is not a substitute for genuine efficacy trials like tenofovir's, where the value of the intervention is exactly what remains unknown.

6. Future Work

The arguments above are largely conceptual, and the empirical question remains open: it is not known how many current trials enrolling people who inject drugs actually meet guidance point 13. A systematic audit of registered trials, coding prevention packages, ancillary care provisions, consent documents, and site-level deviations against that standard, would establish whether the conditions proposed here describe a reform or merely current practice. Until that audit exists, the field is about a problem whose size is not measured.

7. References

[1]Singh JA, Mills EJ. The abandoned trials of pre-exposure prophylaxis for HIV: what went wrong? PLoS Med. 2005;2(9):e234. doi:10.1371/journal.pmed.0020234

[2]Alcorn K. Thai tenofovir trial runs into trouble after ethics protests from drug users. aidsmap. Published March 24, 2005. Accessed September 4, 2026. https://www.aidsmap.com/news/mar-2005/thai-tenofovir-trial-runs-trouble-after-ethics-protests-drug-users

[3]Choopanya K, Martin M, Suntharasamai P, et al. Antiretroviral prophylaxis for HIV infection in injecting drug users in Bangkok, Thailand (the Bangkok Tenofovir Study): a randomised, double-blind, placebo-controlled phase 3 trial. Lancet. 2013;381(9883):2083-2090. doi:10.1016/S0140-6736(13)61127-7

[4]World Medical Association. WMA Declaration of Helsinki: ethical principles for medical research involving human participants. Updated October 2024. Accessed September 4, 2026. https://www.wma.net/policies-post/wma-declaration-of-helsinki/

[5]Emanuel EJ, Wendler D, Grady C. What makes clinical research ethical? JAMA. 2000;283(20):2701-2711. doi:10.1001/jama.283.20.2701

[6]Freedman B. Equipoise and the ethics of clinical research. N Engl J Med. 1987;317(3):141-145. doi:10.1056/NEJM198707163170304

[7]MacArthur GJ, Minozzi S, Martin N, et al. Opiate substitution treatment and HIV transmission in people who inject drugs: systematic review and meta-analysis. BMJ. 2012;345:e5945. doi:10.1136/bmj.e5945

[8]Joint United Nations Programme on HIV/AIDS, World Health Organization. Ethical Considerations in Biomedical HIV Prevention Trials. Additional guidance point added 2012. UNAIDS; 2012. Accessed September 4, 2026. https://www.unaids.org/sites/default/files/media_asset/jc1399_ethical_considerations_en_0.pdf

[9]Richardson HS, Belsky L. The ancillary-care responsibilities of medical researchers: an ethical framework for thinking about the clinical care that researchers owe their subjects. Hastings Cent Rep. 2004;34(1):25-33. doi:10.2307/3528248

[10]DeBeck K, Cheng T, Montaner JS, et al. HIV and the criminalisation of drug use among people who inject drugs: a systematic review. Lancet HIV. 2017;4(8):e357-e374. doi:10.1016/S2352-3018(17)30073-5

[11]Miller WC, Hoffman IF, Hanscom BS, et al. A scalable, integrated intervention to engage people who inject drugs in HIV care and medication-assisted treatment (HPTN 074): a randomised, controlled phase 3 feasibility and efficacy study. Lancet. 2018;392(10149):747-759. doi:10.1016/S0140-6736(18)31487-9

[12]Haire BG. Because we can: clashes of perspective over researcher obligation in the failed PrEP trials. Dev World Bioeth.

2011;11(2):63-74.
[13]Philpott S, Heise L, McGrory E, Paxton L, Hankins C. The challenge of defining standards of prevention in HIV prevention trials. J Med Ethics. 2011;37(4):244-248.
[14]Dawson L, Zwerski S. Clinical trial design for HIV prevention research: determining standards of prevention. Bioethics. Published online 2015. doi:10.1111/bioe.12113
[15]Dawson L, Strathdee SA, London AJ, et al. Addressing ethical challenges in HIV prevention research with people who inject drugs. J Med Ethics. Published online 2016. https://pubmed.ncbi.nlm.nih.gov/27114469/
[16]Aspinall EJ, Nambiar D, Goldberg DJ, et al. Are needle and syringe programmes associated with a reduction in HIV transmission among people who inject drugs: a systematic review and meta-analysis. Int J Epidemiol. 2014;43(1):235-248. doi:10.1093/ije/dyt243
[17]Larney S, Peacock A, Leung J, et al. Global, regional, and country-level coverage of interventions to prevent and manage HIV and hepatitis C among people who inject drugs: a systematic review. Lancet Glob Health. 2017;5(12):e1208-e1220. doi:10.1016/S2214-109X(17)30373-X
[18]Council for International Organizations of Medical Sciences, World Health Organization. International Ethical Guidelines for Health-Related Research Involving Humans. 4th ed. CIOMS; 2016. Accessed September 4, 2026. https://cioms.ch/publications/product/international-ethical-guidelines-for-health-related-research-involving-humans/
[19]Wertheimer A. Rethinking the Ethics of Clinical Research: Widening the Lens. Oxford University Press; 2011.

Kiruna, 2021: A Case Study into Moral Hazard and a Test for Moral Hazard Objections to Research

Sean Minjune Park[1*], Luke Zhang[2*]
[1]*Chattahoochee High School, Johns Creek, GA 30022,* [2]*Southfield Christian School, Southfield, MI 48034*

Abstract

Some research is restricted because of what it might release. In February 2021, the Saami Council and three environmental organizations demanded cancelation of a balloon flight (SCoPEx) near Kiruna because of what it might reveal.[1] The flight was an equipment test, would have released no material, and was cleared by independent reviews,[2],[3] but the Saami objected that stratospheric aerosol injection constitutes a moral hazard;[1],[4] essentially an objection to producing knowledge, a recurring theme in virology and nuclear weapons research. This paper examines whether knowledge makes action more likely, whether that justifies restricting inquiry, and proposes a three-part framework for evaluating the moral hazard objections to research. Existing frameworks like the Fink criteria[5] and HHS P3CO[6] ask only whether knowledge confers capability, necessary but insufficient, since neither addresses moral hazard.

We derive the framework from proportionality analysis, the method widely used to evaluate restrictions on recognized freedoms, including the freedom of scientific research. Prohibition is warranted only where three conditions hold: (1) producing the knowledge reduces effort on the problem it addresses; (2) restricting it prevents the work altogether; (3) research enables the harmful act. SCoPEx satisfies none: the evidence does not show that awareness of solar geoengineering reduces support for cutting emissions,[7] the cancellation was followed by research continuing among state and private actors rather than stopping,[2] and describing a danger of stratospheric aerosol injection is separate from outlining how to deploy it.[8] Because the analysis is qualitative and rests on a single case, we present the test as a structured framework to be tested against further cases rather than a settled rule, and outline the empirical work needed to do so.

1. Introduction

1.1 Restriction of inquiry as a recurring problem

Research is oftentimes stopped before it is carried out due to hazards within the experiment itself and its design. However, a different class of cancellations has started to form: experiments halted because of what they are expected to lead to. Governments have at times classified findings of such work, and funders as well as institutions have canceled projects because the results could lead to potentially dangerous consequences.

The core question at the heart of restricting research in the past was, does the work confer a capability that someone dangerous did not previously have? Existing review frameworks like the Fink criteria[5] and HHS P3CO[6] only examine whether knowledge of a specific technique or mechanism can make someone else capable of it. This is necessary but insufficient to cover issues like the one at Kiruna, which center on moral hazard objections.

A second type of objection is that the effects of the research (production of some new knowledge) will change what people believe about the future, and they will act differently as a result. In climate policy, this argument is called mitigation deterrence.[4] If governments and the public know that solar geoengineering and

atmospheric modification are being researched, they will treat emissions as less urgent and will cut them less. Essentially, this is an objection to the hazards that arise when conducting research into a remedy for a problem, which will reduce effort into avoiding the problem in the first place.

Moral hazard objections are difficult to judge because no established procedure exists for deciding when they are valid. Without such a procedure, these objections are often resolved inconsistently, according to which party is most influential in a given dispute rather than according to a shared standard.

1.2 Case and contribution

This paper explores a balloon flight scheduled near Kiruna, Sweden, in 2021 and canceled that March after the Saami Council and three environmental organizations demanded that the project be shut down.[1] The objection was not that the flight was unsafe but that the project would explore a technology that they believed had dangerous consequences. Kiruna is therefore an example of moral hazard objections.

This paper argues that a moral hazard objection justifies restricting research only where three conditions hold together. The first is displacement: producing the knowledge reduces effort on the problem it addresses. The objection is at bottom a claim about behavior, so the condition is met only where the evidence shows that people do less about a problem once they know a remedy is being studied. The second is containment: restricting the research prevents the work altogether rather than relocating it. A prohibition that only moves cheap and widely proposed work toward actors who disclose less does not achieve what the objection asks of it. The third is enablement: research enables the harmful act rather than only describing it. Measuring how a proposed intervention would behave is not the same as supplying the means to carry it out, and this is the point at which the existing capability frameworks do genuine work.

The three conditions are derived from proportionality analysis, the standard method in international human rights law and in many constitutional systems for deciding when a recognized freedom may be limited, applied to the specific claim that a moral hazard objection makes. Because a restriction must pass every step of that analysis, the conditions are joint, meaning that all three must hold to support the prohibition of research on moral hazard grounds. An objection that fails any one of them may still identify a real problem and justify remedies short of prohibition. Section 4 develops each condition in full, and Section 5 applies them to SCoPEx, which satisfies none.

1.3 Scope and limits

There are six limits to state. First, the test governs prohibition, not governance in general. There are other objections to consider, moral hazard simply being one of them. Second, the test addresses moral hazard objections specifically. Where research does confer a dangerous capability, the existing frameworks apply. Third, the paper takes no position on whether stratospheric aerosol injection should ever be deployed. Fourth, the first condition surrounding displacement is empirical and requires outside evidence. Fifth, the framework is developed and applied through a single case. Kiruna was chosen because it presents the moral hazard objection in an unusually clean form, with no capability to confound it, which makes it well suited to showing that existing frameworks leave a gap. One case is sufficient to show that such a gap exists, but not to show that the proposed test fills it correctly across other kinds of research. Claims about other domains in Section 7 are therefore offered as hypotheses rather than findings. Sixth, the analysis is qualitative. It relies on documentary sources and

published studies rather than original data, and each condition is assessed as met or unmet rather than measured. Section 7.3 sets out these limitations in more detail, and Section 7.4 describes the research that would address them.

1.4 Approach

The paper combines conceptual analysis with a documentary case study. The primary sources for the case are the objectors' open letter, the final report of the SCoPEx Advisory Committee and its appendices, and the SCoPEx research plan. Secondary sources are the published empirical literature on the mitigation deterrence and reporting on subsequent activity in the field. Much of the case record is drawn from the Advisory Committee's report, which is the most complete documentation available but was produced within the institution whose project was at issue; where possible, we note independent sources.

2. The Kiruna Case

2.1 Flight description

SCoPEx was designed in two phases, neither of which was ultimately carried out. The first was an equipment test: an unmanned balloon would lift an instrument gondola to roughly 20 km above the Esrange Space Center near Kiruna, Sweden, to establish whether such an instrument could serve as a mechanism for observing the results of an SAI (stratospheric aerosol injection), the second phase.[3] Although never scheduled, research plans indicated the release of 100-2,000 g of calcium carbonate or sulfate into the atmosphere.[3] Phase 1 was physically incapable of performing an assessment of SAI or solar geoengineering of any kind due to it not carrying the aerosol hardware necessary for such an experiment.[8] A standard engineering review found the gondola safe for launch,[2] while an independent legal review found no Swedish or international bar to the flight.[3] The objection therefore could not rest on the flight's technical capabilities. It rested on what conducting it would mean for the trajectory of the program and climate research as a whole.

2.2 Moral Hazard

The letter of 24 February 2021 makes a key demand: "SAI is a technology that entails risks of catastrophic consequences, including the impact of uncontrolled termination and irreversible sociopolitical effects that could compromise the world's necessary efforts to achieve zero-carbon societies... [and] must not be advanced in the absence of full, global consensus on its acceptability."[1]

The letter does acknowledge the fact that Phase 1 would have released nothing, but argues that cancellation was nonetheless necessary because the flight cannot be separated from the larger project it is meant to enable.[1] This objection brings in the question of moral hazard: even before the effects of releasing materials are known, the mere act of conducting research into solar geoengineering could change governments' and the public's thoughts about the technology and, as a consequence, how seriously they pursue emissions reductions. Therefore, the objection to the research is to what the research will cause people to believe and do.

2.3 Results

The SCoPEx Advisory Committee replied on 2 March, acknowledging concerns surrounding the flight, including moral hazard.[9] On 31 March, it called for the flight to be delayed, and Harvard and the Swedish Space Corporation announced the same day that it would not proceed.[2,10] The ground given was procedural. The research team suspended SCoPEx in 2023, and Harvard abandoned it in 2024; no aerosol was released.[2,11]

3. Existing Frameworks

3.1 Capability Conference

Every framework we have identified for stopping research before it happens asks the same question: does this work teach someone how to do something dangerous that they could not do before?

The clearest case is the 2004 Fink report, written for the U.S. government after the anthrax letters. It names seven kinds of biological experiments that have to be reviewed before anyone performs them.[5] The HHS P3CO framework asks the same question in a narrower domain, requiring department-level review of proposed research that is reasonably anticipated to create, transfer, or use an enhanced potential pandemic pathogen, and turning on whether the work would produce an agent capable of causing harm that could not be caused before.[6] Additionally, under the Atomic Energy Act, anything describing how to design a weapon or make fissile material is classified the instant it exists.[12]

3.2 Ability vs. Belief

Using any previous frameworks at the SCoPEx flight produces no meaningful result. The balloon carried no aerosol release system, and the design was published in 2014, containing no classified materials.[8,13] The Fink framework applied to Kiruna finds no dangerous technique being propagated, while applying the Atomic Energy Act finds nothing secret to classify.

However, objectors to the flight did not claim the flight would do any of the previous items that would serve as grounds for typically restricting research. Rather, they claimed the flight would change what governments and the public expect the future to hold and that they would cut emissions less as a result.[1] In this case, the harm alleged runs through belief rather than ability, and belief behaves differently in three ways: agent, damage, and remedy.

3.3 Structural Differences between Capability and Belief

These three differences are structural rather than incidental, and each one bears on why a framework built for capability cannot be applied to a case like Kiruna.

The first difference between ability and belief is the agent, meaning whoever the rule is supposed to act on. Work with a legitimate and harmful purpose is controlled by denying a specific bad actor something they need. However, there is no bad actor at Kiruna. The people whose behavior is supposed to change are governments and ordinary publics, reacting reasonably to what they believe is coming.

The second is the damage. Capability is permanent: it cannot be unlearned, which is why these frameworks exist in the first place. On the other hand, a belief tends to move with the evidence and can be corrected by additional information.

The final difference is the remedy. Secrecy is often what each framework suggests as the solution, often working because in those cases, the disclosure is the harm. However, at Kiruna and similar cases, disclosure is a double-edged sword. If research showed that stratospheric aerosol injection is weaker, riskier, or more expensive than its advocates hope, publishing that would strengthen the case for cutting emissions, not weaken it. We are not aware of any existing review framework that governs harm that travels through belief.

4. The Three-Part Test

4.1 Structure of the test

Section 3 showed that existing frameworks ask only whether research confers a capability, and that this question returns nothing

useful when the alleged mechanism of harm is belief. This section (4) presents a framework for evaluating the harm of research creating belief. A moral hazard objection justifies restricting research only if all three of the following conditions hold:

Displacement. Does producing this knowledge reduce effort on the problem it addresses?

Containment. Would restricting this research prevent the work or relocate it?

Enablement. Does this research enable the harmful act or only describe it?

The conditions are joint (i.e., all three must hold to support the prohibition of research on moral hazard grounds). Section 4.2 derives the conditions from established principles, Section 4.3 explains why they are preferred to alternative tests, Sections 4.4 to 4.6 develop each condition, and Section 5 applies the test to Kiruna.

4.2 Derivation from established principles

The three conditions follow from applying an established method for evaluating restrictions on a recognized freedom to the specific causal claim that a moral hazard objection makes.

The starting point is that scientific inquiry is a recognized freedom. Article 15(3) of the International Covenant on Economic, Social and Cultural Rights commits states to "respect the freedom indispensable for scientific research," and the UNESCO Recommendation on Science and Scientific Researchers affirms the freedom of researchers to pursue and communicate their work.[14,15] This does not make research immune to restriction, but it places the burden of justification on the party seeking restriction. The same allocation follows from Mill's harm principle, under which liberty may be limited only to prevent harm to others.[16]

Where a recognized freedom is limited, the standard method of evaluation in international human rights law and in many constitutional systems is proportionality analysis.[17,18] In its common form, a restriction is justified only if (a) it serves a legitimate aim grounded in a real risk of harm; (b) it is suitable, meaning that it actually advances that aim; (c) it is necessary, meaning that no less restrictive measure would achieve the aim as well; and (d) its benefits outweigh its costs in the particular case. The Siracusa Principles, which interpret the limitation clauses of the International Covenant on Civil and Political Rights, state similar requirements: a limitation must respond to a pressing need, pursue a legitimate aim, be proportionate to that aim, and use no more restrictive means than are required.[19]

A moral hazard objection has a specific causal structure. It claims that a piece of research will produce knowledge, that the knowledge will change behavior in a harmful way, whether by reducing mitigation effort or by making an eventual harmful deployment more likely, and that prohibiting the research will prevent this. Applying the steps of proportionality analysis to that structure yields the three conditions.

Legitimate aim yields displacement. A restriction can only serve the aim of preventing harm if the harm is real. Because the harm alleged by a moral hazard objection runs through behavior, whether it is real is an empirical question: does producing the knowledge in fact reduce effort on the problem?

Suitability yields containment. A restriction that does not advance its aim fails suitability, however serious the aim. For a prohibition, the aim is advanced only if the work does not simply take place elsewhere.

Attribution of harm yields enablement. Proportionality presupposes that the restricted conduct is what produces the harm. Law and moral philosophy generally attribute

greater responsibility for a harm to those who supply the means of bringing it about than to those who supply information that others may act on through their own free and informed choices.[20] Bostrom's typology of information hazards similarly distinguishes information that supplies a template for causing harm from information that merely draws attention to a possibility.[21] Enablement applies this distinction to research.

The remaining steps, necessity and balancing, do not add further conditions for prohibition. They govern what should happen when an objection fails: if a less restrictive measure would address the concern, prohibition is not necessary, and Section 6 develops those measures. The test is therefore best understood as the threshold that prohibition, the most restrictive available measure, must pass before a proportionality review can support it at all. Because a restriction must pass every step of that review, the three conditions are joint rather than alternatives.

4.3 Why these conditions and not others

A reasonable challenge is that other tests could be proposed that look equally natural. The derivation in Section 4.2 answers this in two ways.

First, each condition is necessary, in the sense that removing any one of them produces results that proportionality analysis would reject. Without displacement, research could be prohibited on the basis of an asserted harm that has never been shown to occur, which fails the requirement of a legitimate aim grounded in a real risk. Without containment, research could be prohibited even when the prohibition prevents none of the alleged harm, and may worsen it by shifting the work to less accountable actors, which fails suitability. Without enablement, research that argues against a technology could be prohibited merely because it concerns that technology, as Section 4.6 illustrates.

Second, the main alternative tests tend either to restate these conditions or to omit one of them. Three alternatives are worth considering.

A precautionary test would restrict research whenever it might contribute to serious harm, even without conclusive evidence. As stated in the Rio Declaration, the precautionary principle holds that a lack of full scientific certainty should not be used as a reason to postpone cost-effective measures to prevent serious environmental degradation.[22] Applied to research, however, the principle cuts in both directions, because research is often what reduces uncertainty about the harm in question, and a prohibition can preserve the very uncertainty that precaution is meant to address. A precautionary approach can also be accommodated within the present test as a lower evidentiary threshold for displacement where potential harms are catastrophic. Section 7.4 treats this as an open question rather than dismissing it.

A consensus test, which the Kiruna letter itself proposes, would allow research to proceed only once there is “full, global consensus on its acceptability.”1 Because any party can withhold consensus, this functions as a veto and reduces to prohibition without requiring any showing of harm. It omits displacement entirely, and where the research can proceed elsewhere, it fails containment.

A cost-benefit test would weigh the expected harms of the research against its expected benefits. The present test is compatible with this approach and can be read as specifying when prohibition has any expected benefit to weigh. If the knowledge does not change behavior, or if the prohibition does not stop the work, the harm averted by prohibiting the research is close to zero, and the costs of

prohibition are borne for little or no gain. Enablement plays a different role on the other side of the ledger. Where research only describes a technology, and particularly where its findings count against deployment, the knowledge it produces is among the main benefits of the work, and prohibiting it forgoes the information most likely to correct an overconfident belief in the remedy.

None of this makes the test irrefutable. A reader who rejects proportionality as the right standard for limiting research, for example in favor of a strong precautionary principle, may arrive at a different test. What the derivation does show is that the test is not arbitrary: it follows from a standard that is already widely used to decide when recognized freedoms may be limited, and disagreement with it can be located in a specific premise rather than in competing intuitions.

4.4 Displacement: Does producing this knowledge reduce effort on the problem it addresses?

This is the factual claim at the heart of the objection. Put concretely: does knowing that a technology is being researched make governments and publics do less about the problem that technology is meant to address? This question tells us how people actually behave, and researchers have already run surveys and experiments testing it. The condition is met when the evidence shows the effect and not met when the studies disagree or when nobody has looked. The burden sits with the objector, who is the one claiming that producing this knowledge will change what people do.

4.5 Containment: Would restricting this research prevent the work, or relocate it?

A ban is justified only if it actually stops the work. If the research is cheap and widely proposed, others may simply carry it out instead. Institutions that accept restrictions are also the ones most likely to face oversight, so canceling their project may only push the work toward actors who disclose less and face less scrutiny.

4.6 Enablement: Does this research enable the harmful act or only describe it?

Describing a phenomenon, estimating its magnitude, or measuring how badly a proposed intervention would perform is not the same as building the intervention. The distinction is noteworthy because a finding that some technology is riskier or weaker than its advocates hope cuts against deployment while still counting as knowledge about it. The condition separates research from deployment on the content of the work rather than on the researchers' stated intentions, which is what objectors are usually unwilling to accept. It is also the point at which the capability frameworks of Section 3 do genuine work: what they ask about ability is what this condition needs answered.

The condition is necessary because the first two do not, by themselves, rule out banning helpful research. Suppose a study finds that a proposed technology works worse than its advocates claim. The finding can satisfy displacement because it changes what people believe, as well as containment. After all, the institution conducting the study can be stopped. Yet banning the study would still make little sense: it gives no one the ability to deploy the technology, and its findings argue against using it. Enablement captures this distinction. The research must make the feared act easier to perform, not merely provide information about it.

5. Applying the Test to SCoPEx

Section 4 outlined a three-part framework to close the gap between cases like

Kiruna and existing frameworks. This section applies that framework to Kiruna and evaluates the implications. The assessments below are qualitative judgements based on the documentary record and the published literature.

5.1 Displacement

Objectors to the experiment claimed that knowledge about solar geoengineering causes governments and the public to reduce their effort on emissions, a claim about behavior, which has been tested directly. Merk and Wagner's 2024 study is the largest such test: a revealed-preference experiment showing climate messages to about 340,000 social media users in the United States to measure whether exposure to information about solar radiation management or carbon dioxide removal changed engagement with a large environmental non-profit's mission relative to baseline climate messaging.[7] They found that the messages had no impact in either direction.

Although the evidence is not incredibly conclusive, as it is drawn from one country on one platform and possibly an unrepresentative sample, the burden of proof still lies with the objector who makes the affirmative claim. The best available test does not find that effect. On the framework's own terms, the condition is not met. This conclusion is provisional and could change as further evidence becomes available, particularly on how governments rather than individuals respond.

5.2 Containment

The cancellation of the experiment in Kiruna did not stop further experiments in the field. SCoPEx was abandoned in 2024, but between the time initial objections were raised and the final cancellation, related activity continued and in some respects grew.[2,11] A private company launched stratospheric balloons carrying sulfate without prior announcement;[23] a marine cloud brightening trial run by a university team was kept closely held to avoid the kind of opposition SCoPEx had attracted;[24] and government and philanthropic funding for the field grew.[25] As noteworthy as the growth in scale of these experiments was, the shift in nature and composition is more concerning. SCoPEx had a published design, external advisory committees on a variety of topics, independent legal boards, and a clear commitment to open publication and full disclosure of funding and results, as well as an explicit exclusion of fossil fuel funders from their study.[2,26] These are the markers of accountable research, and they exist only in institutions exposed to some sort of authority or pressure. Actors who disclose nothing are not subject to that pressure.

These events do not show that the cancellation caused the shift. The field might have developed in much the same way had SCoPEx flown, and that counterfactual cannot be observed. What is shown is that the cancellation did not prevent the work, which is what containment requires. The condition is therefore not met at Kiruna, and the available evidence is consistent with the concern that prohibition can move work toward less accountable actors.

5.3 Enablement

The SCoPEx flight was physically incapable of conducting an assessment of solar geoengineering because it lacked aerosol release hardware.[8]

Its design was published in 2014, containing no restricted components or mechanisms; even the second phase, which was never scheduled, would have released between 100 and 2,000 grams of calcium carbonate or sulfate, a quantity chosen to be observable rather than climatically significant and orders of magnitude below anything that could affect the

atmosphere at scale.[3,13]

The SCoPEx program would have produced a measurement of how such aerosols behave in the stratosphere, which is information about the intervention rather than a means of carrying it out. This is precisely the distinction the third condition draws, so the condition is not met.

5.4 Verdict

On the available evidence, no condition is satisfied. Displacement is not supported by the direct evidence that exists; containment fails because the work continued elsewhere, including among actors who disclose less; enablement fails due to the nature of the flight itself. The moral hazard objection does not justify stopping the flight at Kiruna.

These judgements are qualitative, and two of them, displacement and containment, rest on a limited evidence base. A different reading of that evidence, or new evidence, could change the result for a given condition. Because the conditions are joint, however, the verdict would change only if all three were shown to hold.

This is not the same as saying the cancellation was wrong. The Advisory Committee's publicly stated ground was procedural: the flight had been scheduled before any serious engagement with the communities near the launch site took place, and the Committee called for a delay pending that engagement.[2] That objection is not refuted by the argument here. It is not a moral hazard objection, it does not depend on any claim about what the research would cause people to believe, and it calls for a different remedy. Section 6 takes up what those remedies are.

6. Remedies Short of Prohibition

Prohibition is the strongest response to an objection, but it is also likely to fail, since its success depends entirely on the work having nowhere else to go. An objection that fails the test of Section 4 may still find a real problem. As a result, work can either continue uninterrupted or completely stop.

6.1 Disclosure and registration

The 2021 National Academies report *Reflecting Sunlight* recommends a public registry of solar geoengineering research, alongside permitting for outdoor experiments, a research code of conduct, and guidance on intellectual property.[27] Registration and permitting are different and are worth separating. A permit is a decision, and someone may refuse it. A registry records that work is being done and discloses everything.

The distinction matters because of what happened at Kiruna. SCoPEx was among the most openly documented projects in its field, with a published design, an external advisory committee, and disclosed funding sources, and that openness allowed objectors to argue against the project. Section 5.2 described that this resulted in the work continuing afterward to proceed more quietly.

A registry is useful because it has no veto. It separates the question of whether an experiment is known from the question of whether it may proceed, and the conflation of those two questions contributed to the outcome at Kiruna. Objections can still be raised, but they should be raised while a design is still open to revision rather than when the only choices are to go ahead or to cancel the experiment.

6.2 Standing and Permission

Where an experiment takes place is a separate question from whether it takes place, and objections often attach to the first while being expressed as objections to the second. Esrange lies within Sápmi, and part of the grievance at Kiruna was that a test bearing on the future of the Arctic had been scheduled on

Saami land without asking the Saami for permission.[1,28] That is an objection to the site, and it can be resolved by moving the experiment at a cost far below cancellation.

The Advisory Committee's own recommendation also states that any change of location would require a new review, because the legal and social conditions differ from place to place.[2] A project that treats its launch site as a logistical detail should expect the site to become the objection.

6.3 Consent and procedural legitimacy

Consultation and consent are different things, and confusion between them contributed to the failure at Kiruna. Consultation began after the flight had been scheduled, which made it look like ratification of a decision that was already taken.[2,29] The Advisory Committee's guidelines were written against exactly this: begin early, use independent facilitators, make the process two-way and deliberative, and do not presuppose what a community's concerns will be.[30]

Consent is most defensible when understood as authority over local siting rather than as an objection to a research question. This gives a community entitlement to refuse it. A veto over whether anyone anywhere may study a subject reduces to prohibition and inherits the containment problem in full. The difficulty is that solar geoengineering's effects would not be local, and no procedure for identifying a globally relevant community exists. Section 7 returns to this.

6.4 Proportionality

The remedies form a rough ordering by cost, which corresponds to the necessity step of the proportionality analysis in Section 4.2. Disclosure and registration are close to costless and are appropriate for any outdoor experiment, whether or not anyone has objected. Siting review is appropriate once a community bears a local burden. Consent processes are warranted where that burden is material or where the site carries specific rights, as at Esrange. Prohibition is reserved for the case in which all three conditions of Section 4 hold.

What went wrong at Kiruna was that the choice presented was to proceed as scheduled or to cancel, and once framed that way, cancellation was the more defensible option. The proportionate response, which was to engage first and to site the test where the burden falls on a community that has agreed to bear it, was available throughout and was in substance what the Advisory Committee asked for.[2]

7. Conclusion

7.1 Findings

Kiruna presents a moral hazard objection in an unusually clean form. The flight would have released nothing, conferred nothing, and violated nothing. The objection was to what the results would cause people to believe. Existing review frameworks cannot be applied to such a case, because they ask whether research confers a capability, and here it does not.

This paper has proposed that a moral hazard objection justifies prohibiting research only where displacement, containment, and enablement all hold. Applied to Kiruna, none appears to hold. The experimental evidence does not show that information about solar geoengineering reduces mitigation engagement. The cancellation did not prevent the research, which continued, partly among actors who disclose less. The flight would have described a proposed intervention without supplying any means of carrying it out. The objection did not support stopping the flight.

Failing the test is not vindication. The procedural objection at Kiruna, which affected

communities that were engaged only after the flight had been scheduled, was well founded, is not rejected by this analysis, and is answerable by remedies short of prohibition.

7.2 Possible extensions beyond geoengineering

Because the test was developed through a single case, its application elsewhere is a hypothesis rather than a finding. The structure may extend wherever a proposed solution is said to reduce the effort put into avoiding the problem it addresses. Research on carbon dioxide removal and on climate adaptation attracts the same objection in a similar manner. Research on long-term nuclear waste disposal has been said to make new reactors easier to authorize. In public health, the wider availability of overdose-reversal drugs has been argued to encourage riskier drug use.

Whether the test gives sensible results in these domains has not been examined here, and these cases differ from Kiruna in ways that matter. In some, the research may be expensive or concentrated in a few institutions, so containment could hold. In others, such as research on enhanced pathogens, enablement is often clearly met, and the capability frameworks of Section 3 already apply. A test that returned "no prohibition" in every case would be suspect, so an important check on this framework is whether it also identifies cases in which restriction is warranted.

If the test does extend, its usefulness would be that it converts a contest of intuitions into three answerable questions, two of which, containment and enablement, can often be addressed from facts already available about the field and about the work itself.

7.3 Limitations

The analysis has several limitations beyond those noted in Section 1.3.

A single case. The framework was developed with Kiruna in view and then applied to Kiruna. This creates a risk that the conditions fit the features of this case rather than the general problem. Kiruna shows that capability-based frameworks leave a gap, but it cannot show that the proposed test fills that gap correctly in other settings.

Qualitative assessment. Each condition was judged met or unmet from documents and published studies, without quantitative thresholds. In practice, the conditions are matters of degree: displacement effects can be small or large. Containment can be partial, and research can make a harmful act somewhat easier without enabling it outright. The binary treatment here simplifies this, and different analysts applying the test to the same case could reasonably disagree.

The evidence base. The displacement judgement rests heavily on an enlarged experiment, conducted with individuals in one country on one platform. The containment judgement rests on a small number of subsequent events drawn from press reports and one advisory report, and it cannot establish what would have happened had the flight proceeded. Much of the case record comes from the SCoPEx Advisory Committee's report, which was produced within the institution whose project was at issue.

Interpretation of the objection. The objection is reconstructed primarily from a single open letter. The objectors' concerns broader than moral hazard, including consent and the use of Saami land, and a reconstruction focused on moral hazard may understate the weight they placed on those concerns. Section 6 addresses some of them, but this paper does not evaluate them in full.

7.4 Open questions and future research

The available studies are concentrated in a small number of wealthy countries and

measure the responses of individuals. The deterrence claim, however, is about governments, and there is almost no direct evidence on how policymakers respond to the existence of geoengineering research. Until there is, the first condition will often be decided by an absence of evidence rather than by evidence of absence.

The allocation of the burden of proof is also contestable. Placing it on the objector follows from the fact that the objector makes the affirmative causal claim, but where the potential harm is large, this may be the wrong choice, and the framework does not supply a threshold that would justify shifting it.

The connecting structure may be too permissive at the extreme. Where displacement is strongly evidenced and containment holds, but the research only describes rather than enables, the test returns no prohibition. That is the correct result in ordinary cases and is what the third condition produces. Whether it remains right when the stakes are catastrophic is a question this framework does not settle.

Beyond these, the scale problem raised in Section 6 remains open. Consent is a solution for local burdens and has no established form for global ones. A technology whose effects would fall on everyone has no community whose agreement could stand for the rest, and until that gap is addressed, procedural legitimacy for research of this kind will remain contested, even if, on this analysis, the moral hazard objection alone does not justify prohibition.

Several lines of research would address these questions and the limitations in Section 7.3.

Operationalizing the conditions. Each condition could be given measurable indicators. Displacement could be assessed from pooled effect sizes across experimental and survey studies of mitigation deterrence, with a stated threshold for a meaningful effect. Containment could be assessed from the cost of and technical barriers to conducting the work, the number of actors able to do it, and records of undisclosed activity before and after a restriction. Enablement could be assessed with a rubric scoring whether research outputs include deployable methods, specification, or materials. A graded version of the test would allow partial satisfaction of a condition to be weighed rather than treated as all-or-nothing.

Comparative case studies. Studies of how policymakers and negotiators respond to information about geoengineering research, for example through expert elicitation, policy simulation exercises, or analysis of policy documents and negotiating positions over time, would address the largest gap in the displacement evidence.

Reliability. Having independent reviewers apply the test to the same set of cases, and measuring their agreement, would show whether the test is specific enough to be applied consistently by different people, which is a precondition for using it in actual research review.

Of these questions, the missing evidence on how governments respond is the most pressing, because displacement carries the empirical weight of the whole test and is at present the condition most often decided by an absence of evidence rather than by evidence. The scale problem is second, since it determines whether the remedies of Section 6 are available at all where the effects of a technology fall on everyone. The allocation of the burden of proof and the permissiveness of the joint structure matter most in the catastrophic tail, and are the questions on which the framework's answer is least secure, though also the ones least often reached in practice.

8. References

[1]Saami Council, Swedish Society for Nature Conservation, Friends of the Earth Sweden, & Greenpeace Sweden. (2021, February 24). Open letter to the SCoPEx Advisory Committee. In S. Jinnah et al. (2024), Final report of the SCoPEx Advisory Committee (Appendix E-3). Harvard Salata Institute. Retrieved September 4, 2026, from https://static1.squarespace.com/static/5dfb35a66f00d54ab0729b75/t/603e2167a9c0b96ffb027c8d/1614684519754/Letter+to+Scopex+Advisory+Committee+24+February.pdf

[2]Jinnah, S., Bedsworth, L., Talati, S., et al. (2024, March 18). Final report of the SCoPEx Advisory Committee. Harvard Salata Institute. Retrieved September 4, 2026, from https://salatainstitute.harvard.edu/wp-content/uploads/2025/04/Final-SCoPEx-AC-Report-With-Appendices.pdf

[3]Setterwalls Advokatbyrå. (2021, February 18). Legal review memorandum to Harvard University. In S. Jinnah et al. (2024), Final report of the SCoPEx Advisory Committee (Appendix B). Harvard Salata Institute.

[4]McLaren, D. (2016). Mitigation deterrence and the "moral hazard" of solar radiation management. Earth's Future, 4(12), 596–602. https://doi.org/10.1002/2016EF000445

[5]National Research Council. (2004). Biotechnology research in an age of terrorism. National Academies Press. Retrieved September 4, 2026, from https://nap.nationalacademies.edu/catalog/10827/

[6]Office of Science and Technology Policy. (2017, January 9). Recommended policy guidance for departmental development of review mechanisms for potential pandemic pathogen care and oversight (P3CO). Executive Office of the President. Retrieved September 4, 2026, from https://osp.od.nih.gov/wp-content/uploads/2017_OSTP_P3CO_Policy_Guidance.pdf

[7]Merk, C., & Wagner, G. (2024). Presenting balanced geoengineering information has little effect on mitigation engagement. Climatic Change, 177(1), 11. https://doi.org/10.1007/s10584-023-03671-5

[8]SCoPEx Research Team. (n.d.). Flight platform test research plan and supporting correspondence. In S. Jinnah et al. (2024), Final report of the SCoPEx Advisory Committee (Appendices C-1 to C-5). Harvard Salata Institute.

[9]SCoPEx Advisory Committee. (2021, March 2). Reply to the Saami Council. In S. Jinnah et al. (2024), Final report of the SCoPEx Advisory Committee (Appendix E-4). Harvard Salata Institute.

[10]Temple, J. (2021, March 31). Geoengineering researchers have halted plans for a balloon launch in Sweden. MIT Technology Review. Retrieved September 4, 2026, from https://www.technologyreview.com/2021/03/31/1021479/

[11]Temple, J. (2024, March 18). Harvard has halted its long-planned atmospheric geoengineering experiment. MIT Technology Review. Retrieved September 4, 2026, from https://www.technologyreview.com/2024/03/18/1089879/

[12]Atomic Energy Act of 1954, 42 U.S.C. §§ 2014(y), 2161–2168.

[13]Keith, D. W., Duren, R., & MacMartin, D. G. (2014). Field experiments on solar geoengineering: Report of a workshop exploring a representative research portfolio. Philosophical Transactions of the Royal Society A, 372(2031), 20140175. https://doi.org/10.1098/rsta.2014.0175

[14]United Nations General Assembly. (1966, December 16). International Covenant on Economic, Social and Cultural Rights (Article 15(3)). Retrieved September 27, 2026, from https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-economic-social-and-cultural-rights
[15]UNESCO. (2017, November 13). Recommendation on science and scientific researchers. Retrieved September 27, 2026, from https://unesdoc.unesco.org/ark:/48223/pf0000263618
[16]Mill, J. S. (1859). On liberty. John W. Parker and Son. Retrieved September 27, 2026, from https://www.gutenberg.org/ebooks/34901
[17]Barak, A. (2012). Proportionality: Constitutional rights and their limitations. Cambridge University Press. https://doi.org/10.1017/CBO9781139035293
[18]Alexy, R. (2002). A theory of constitutional rights (J. Rivers, Trans.). Oxford University Press. Retrieved September 27, 2026, from https://global.oup.com/academic/product/a-theory-of-constitutional-rights-9780198258216
[19]United Nations Commission on Human Rights. (1984, September 28). The Siracusa principles on the limitation and derogation provisions in the International Covenant on Civil and Political Rights (UN Doc. E/CN.4/1985/4). Retrieved September 27, 2026, from https://www.refworld.org/legal/resolution/unchr/1984/en/57200
[20]Hart, H. L. A., & Honoré, T. (1985). Causation in the law (2nd ed.). Clarendon Press. https://doi.org/10.1093/acprof:oso/9780198254744.001.0001
[21]Bostrom, N. (2011). Information hazards: A typology of potential harms from knowledge. Review of Contemporary Philosophy, 10, 44–79. Retrieved September 27, 2026, from https://nickbostrom.com/information-hazards.pdf
[22]United Nations. (1992, June). Rio declaration on environment and development (UN Doc. A/CONF.151/26, Vol. I, Annex I, Principle 15). Retrieved September 27, 2026, from https://sdgs.un.org/documents/aconf15126-vol-i-rio-declaration-19576
[23]Temple, J. (2022, December 24). A startup says it's begun releasing particles into the atmosphere, in an effort to tweak the climate. MIT Technology Review. Retrieved September 4, 2026, from https://www.technologyreview.com/2022/12/24/1066041/
[24]Sumagaysay, L. (2024, June 5). Alameda officials stop first-of-kind geoengineering experiment. San Francisco Chronicle. Retrieved September 4, 2026, from https://www.sfchronicle.com/climate/article/geoengineering-alameda-study-stopped-19453924.php
[25]Carbon Brief. (2025). Factcheck: How the UK is – and is not – studying solar geoengineering. Retrieved September 4, 2026, from https://www.carbonbrief.org/factcheck-how-the-uk-is-and-is-not-studying-solar-geoengineering
[26]Keith, D. W. (2020, June 10). Statement regarding financials and conflict of interest. In S. Jinnah et al. (2024), Final report of the SCoPEx Advisory Committee (Appendix A-2). Harvard Salata Institute.
[27]National Academies of Sciences, Engineering, and Medicine. (2021). Reflecting sunlight: Recommendations for solar geoengineering research and research governance. National Academies Press. https://doi.org/10.17226/25762
[28]Oksanen, A. A. (2023). Dimming the midnight sun? Implications of the Sámi Council's intervention against the SCoPEx project. Frontiers in Climate, 5, 994193.

https://doi.org/10.3389/fclim.2023.994193
[29]SCoPEx Advisory Committee. (2021, January 8). Proposed engagement process for SCoPEx (Final version). In S. Jinnah et al. (2024), Final report of the SCoPEx Advisory Committee (Appendix E-1). Harvard Salata Institute.
[30]SCoPEx Advisory Committee. (2022, October 26). Guidance on local engagement. In S. Jinnah et al. (2024), Final report of the SCoPEx Advisory Committee (Appendix E-2). Harvard Salata Institute.

Journal of Ethics in Scientific and Technological Innovations

Seeding the Oceans: The Divide Between Marine Conservation and Progressive Science

Alex Fu[1*], Leroy Jing[1*], Roy Lin[1*]
[1]Union County Magnet High School, Scotch Plains, NJ 07076, [2]William Fremd High School, Palatine, IL 60067, [3]Munster High School, Munster, IN 46321

Abstract

Human activities are causing our oceans to face many critical ecological issues. They absorb 90% of excess global heat and 30% of carbon dioxide. This changing underwater climate is killing wildlife like coral and shellfish at an unprecedented rate. Combined with issues like plastic pollution, overfishing, and habitat loss, we must find a way to reverse these compositional changes to the oceans in order to rescue those vulnerable ecosystems.

Ocean iron fertilization is a proposed geoengineering technique aimed at preventing the effects of climate change on the oceans. It involves depositing large amounts of iron into the ocean in order to create massive phytoplankton blooms. They absorb large amounts of carbon dioxide and leave it on the ocean floor when they die. However, this solution leaves ethical concerns in its side effects, especially in marine conservation. It can create toxic algal blooms and disrupt food webs. Ultimately, we will explore whether or not slowing the warming of our planet is worth the cost of using our oceans as an experimental flask.

1. Introduction

Iron fertilization of the ocean has been suggested as a geoengineering technique to slow climate change impacts on the oceans.[1] This concept addresses a well-known limitation in ocean chemistry: that large areas of open ocean, such as the Southern Ocean, have extremely low concentrations of dissolved iron.[2] This constraint restricts the phytoplankton growth to a level lower than it could be with the available nutrients.[3] Adding iron to these waters would get phytoplankton blooms big enough to be visible from space.[3] Additionally, as these organisms photosynthesize, they take in carbon dioxide from the air.[4] When these organisms die, they sink to the ocean floor with the carbon they captured, holding it there for decades or even longer.[4]

This approach isn't hypothetical, with many studies already conducted.[5] Between 1993 and 2009, thirteen field experiments tested if these iron additions could reliably capture carbon and sink it to the deep ocean.[6] Only one of those thirteen showed a measurable increase in carbon going to depth. These mixed results, along with concerns over unintended ecological impacts, caused field research to be stopped for over a decade.[7] International regulation was also implemented, with the London Convention and Protocol seeing ocean fertilization as a cautionary framework.[8]

However, that pause is ending. A new generation of ocean iron fertilization trials, bigger and longer than before and monitored more rigorously, are being called for by researchers.[9] In 2026, a large modeling study simulated sixty years of fertilization in ten ocean regions using a detailed global ocean biogeochemistry model.[10] The results were not uniform, and there was no predictable way that carbon removal potential and ecological cost went together.[10] Some areas saw substantial carbon uptake with little disruption, but others,

particularly in the equatorial Pacific, showed that fertilization can deplete the surrounding water of nutrients needed elsewhere, with repercussions up the food chain to zooplankton populations far from the fertilized patch.[10]

This is the heart of the debate with using ocean iron fertilization as a strategy to fight climate change.[11] The same process that could extract carbon from the atmosphere also manipulates living ecosystems, with the effects not fully mapped out.[11] The ethical question this paper asks is, considering this trade-off, is the benefit to the climate worth treating the ocean as a giant test bed for a technology we don't fully understand?[12]

2. Background and Case Studies

2.1 Literature Review

OIF has been extensively explored as a potential mechanism for marine CO2 removal (mCDR).[5] Past literature shows a change from small-scale, short term artificial experiments to comprehensive natural blooms and long term predictive modeling.[10]

Although earlier studies were able to successfully demonstrate that adding dissolved iron to high nutrient, low chlorophyll (HNLC) regions stimulates phytoplankton growth, they were limited in scale and duration.[14] A comprehensive review of thirteen artificial ocean iron fertilization (aOIF) experiments conducted between 1990 and 2009[5] emphasized that these trials only lasted weeks, which was too brief to adequately capture the full life cycle of a phytoplankton bloom and its carbon export.[5] For example, the Southern Ocean European Iron Fertilization (EIFEX) was the only trial to record significant carbon export to deep layers.[6] This was primarily due to its long observation period of 39 days which was long enough to witness the formation of rapidly sinking aggregates.[6] On the other hand, past field experiments were intentionally not designed to quantify the durability or ability of the iron addition as a mCDR strategy.[14] In order to resolve this, future field studies should span significantly larger scales of approximately 1,000 km2 and prolong periods of three to six months to track the complete transition of bloom development to decay and then eventual return to baseline conditions.[14]

Moving from past methodologies requires both updated tracking techniques and rigorous monitoring of carbon pathways.[14] Earlier trials relied on sulfur hexafluoride, SF6, alongside drifting buoys for about two weeks due to air-sea gas exchange.[5] Future studies should be pairing autonomous vehicles and optical sensors with buoyant sediment to accurately measure the carbon pump.[14] The main goal of these trials is to verify that organic carbon which sinks deep enough, around 500 meters, is able to remain sequestered from the atmosphere for at least a century.[14] The Northeast Pacific Ocean has been identified as a highly promising site for such trials, due to its low eddy kinetic energy and proximity to established baseline monitoring stations.[14]

Importantly, any future aOIF deployment should be weighed against unintended ecological and biogeochemical consequences.[15] Both historical and recent analyses emphasize the need to monitor secondary greenhouse gas emissions, like N2O and CH4, generated by the bacterial remineralization of sinking organic matter.[15] Furthermore, shifts in phytoplankton structures can lead to increases in toxic species.[16] Heavily silicified diatoms like Pseudo-nitzschia can produce the neurotoxin domoic acid, though significant accumulation impacting higher trophic levels was not observed in open-ocean trials.[16] To navigate these risks, researchers are relying on high-resolution models to simulate nutrient robbing, dissolved, oxygen depletion, and shifts in marine food webs.[17] These simulations demonstrate stark disparities. Interventions in the Southern Ocean show higher

carbon sequestration and ecological resilience, while applications near the equatorial Pacific risk expanding oxygen-minimum zones and severely depleting macrozooplankton biomass across downstream areas.[10]

These findings highlight the necessity of strictly defined criteria and governance before scaling OIF as a viable climate intervention.

2.2 Scientific Background

The foundation of ocean iron fertilization is based in the "iron hypothesis", famously defended by oceanographer John Martin in the late 1980s and formally published in 1990.[18] Prior to Martin's discoveries, the scientific community thought that marine algae was mostly distributed by the availability of macronutrients like nitrogen and phosphorus. Martin observed, however, that specific parts of the ocean, such as the Southern Ocean, Equatorial Pacific, and Subarctic Pacific, remained seemingly unproductive despite having high concentrations of essential macronutrients.[19] He hypothesized that phytoplankton growth in these high nutrient, low chlorophyll (HNLC) zones was restricted by a lack of the trace micronutrient, iron, which is required for cellular biochemistry and photosynthesis.[19] The famous quote, "Give me half a tanker of iron, and I'll give you an ice age" postulated that introducing iron could spark massive algal blooms, which would remove large amounts of carbon dioxide from the atmosphere.[19]

This climate intervention relies heavily on the deliberate amplification of the marine biological carbon pump.[20] When iron is introduced to HNLC waters, it rapidly stimulates the photosynthetic activity of local phytoplankton populations, especially larger organisms like diatoms.[3] As they bloom, they consume inorganic carbon from the waters, converting it into organic carbohydrates to build their biomass.[4] When these organisms eventually die, or are consumed by zooplankton, this organic carbon begins to sink out of the sunlit euphotic zone. If this organic material descends deep enough into the ocean before being remineralized by bacteria, the carbon remains effectively sequestered from the atmosphere for centuries, establishing a vertical carbon gradient that drives further atmospheric drawdown.[4]

Martin's theory was validated shortly after his death, sparking series of groundbreaking small scale field experiments throughout the 1990s and 2000s.[5] Researchers utilized ultra-clean trace metal sampling techniques to confirm that baseline iron concentration in HNLC waters were low and that artificial additions would trigger massive localized phytoplankton blooms.[5] Over a dozen open-ocean experiments successfully demonstrated that iron enrichment leads to significant localized drawdowns of both macronutrients and dissolved inorganic carbon, contrasting traditional models. While the efficiency of these induced blooms proved that trace metals increased oceanic productivity, ongoing research continues to look at the long-term sequestration efficiency of the pump and broader ramifications of nutrient manipulation.[14]

2.3 Case Study - Benefits

OIF has clear benefits as a natural pathway to carbon capture, which will remove greenhouse gases from the atmosphere and slow down global warming.[3] It offers a low cost alternative to traditional land based carbon capture techniques as only trace amounts of iron are required, which is already an inexpensive material to begin with.[14] Since it leverages the ocean's biology, it could be a potentially scalable form of mCDR. Introducing small amounts of iron to high-nutrient, low-chlorophyll regions, scientists can analyze the multiplier effect of phytoplankton growth, as a single atom of iron can lead to tens of

thousands of carbon atoms.[20] Beyond its potential, stimulating climate blooms can have multiplying positive effects on the marine food web.[3] Increase production provides a food source for zooplankton, which in response supports larger predators. This supports marine biodiversity in historically barren oceanic zones.

In addition, the phytoplankton with the carbon dioxide naturally sinks to the bottom of the ocean and stays there, making OIF a viable long term solution for locking away atmospheric carbon.[4] As the diatoms created by stimulated phytoplankton blooms die or are consumed by zooplankton, they form dense compactions of organic carbon referred to as "marine snow".[3] Because of the added weight of their silica-based shells, they have increased likelihood of sinking rapidly through the water column. This rapid vertical descent is vital for effect mCDR as it reduces the amount of material that is broken down and comes back into the surface waters by bacteria before reaching isolation.[4]

For OIF to function as a durable climate intervention, this sinking organic matter must bypass shallow remineralization and reach specific oceanic depths where it can be securely locked away.[14] It is generally agreed that for carbon to effectively sequestered, it must sink below a threshold, often around 1,000 meters, into the deep ocean or the sea floor. At these depths, the biological carbon is isolated from atmospheric exchange, with potential to remain in deep water nutrient cycles and sediments for hundreds to even thousands of years.[6] Recent biogeochemical modeling shows that if the average depth of carbon remineralization is pushed deeper, the gradients of inorganic carbon could lead to a substantial decrease in global atmospheric CO2 levels.[10] By accelerating the volume of biomass falling to those layers, iron fertilization actively strengthens the biological pump. This maximizes the fraction of carbon that is permanently removed from the active climate system.[20]

This is supported by paleoclimatological records which demonstrate that nature has used this for millions of years.[21] Samples extracted from Vostok station in Antarctica, stretches back over 400,000 years and shows a historical correlation between periods of high iron dust deposition, massive oceanic phytoplankton activity, and significantly lowered atmospheric CO2 levels.[21] During these fertilization events, tens of billions of tons of carbon was drawn out of the atmosphere and deposited into the ocean. This drove global temperatures down and contributed to the onset of glacial periods. By biomimicking these naturally occurring oceanic phenomena, aOIF wants to replicate this historical carbon export. It could potentially act as a scalable method to accelerate deep ocean carbon storage and buy humanity time to transition from fossil fuels.[21]

2.4 Case Study - Ethical Dilemmas

Evaluating a novel geoengineering technology like OIF requires balancing the precautionary principle against proportionality. Are the ecological uncertainties of OIF severe enough to prohibit deployment? Or is the climate crisis urgent enough to justify acting despite the risks?

From a precautionary and ecocentric perspective, large-scale OIF treats living marine ecosystems as mere tools for human carbon management. Early trials produced mixed results. This casts doubt on whether manipulating the base of the marine food web will even stabilize the climate.[7] Rapidly inflating phytoplankton populations carries serious ecological risks. It can trigger toxic algal blooms, expand mid-water dead zones, and release potent greenhouse gases like nitrous oxide and methane as bacteria break down sinking matter.[15,16] Open-ocean interventions cannot be contained or recalled once deployed. Full-scale OIF uses the biosphere as an experimental flask with no way to reverse

unintended damage.[22] Whether we value marine life for its own sake or for the resources it provides to humans, restructuring ocean food webs carries heavy moral weight.

Deploying OIF too soon also creates social and political risks, mainly moral hazard (or mitigation deterrence) and intergenerational injustice. Promising a cheap techno-fix can weaken the political will to cut industrial emissions, which are the root cause of climate change.[23] Furthermore, using ocean fertilization to offset ongoing pollution rather than cutting emissions traps future generations. They would inherit a high-emission economy alongside the burden of maintaining marine interventions forever. If society later halts OIF due to toxic side effects or war, atmospheric carbon levels would surge alongside unchecked industrial emissions. Future populations would be locked into rapid warming with no easy alternatives.[23]

Yet refusing to act carries its own massive human cost. The planet is hurtling toward a 2.0°C temperature increase. A 2026 Oxford study projects that extreme heat will make large regions unlivable for 3.79 billion people by 2050. That is roughly half of the world's population.[25] Proponents of proportionality argue that when billions face severe climate disruption, refusing to explore scalable carbon removal is itself an ethical failure.

This clash between caution and urgency brings up two final concepts: distributive justice and procedural justice. Neither the harms of climate change nor the ecological costs of OIF hit the world equally. Because the high seas belong to no single nation, a wealthy country could fertilize distant international waters to meet its own climate targets. Meanwhile, vulnerable coastal nations would suffer the ecological fallout. The 2026 global modeling study proved this risk is real. Fertilizing the equatorial Pacific can rob downstream waters of vital nutrients and collapse zooplankton populations thousands of kilometers away.[10] In that scenario, industrial powers get the climate benefit while Global South communities lose the fisheries they rely on for food and income.[24] A lack of procedural justice makes this unfair distribution even worse. No binding international body exists to get informed consent from affected nations or to police cross-border ecological damage.[8,24]

3. Discussion

The case against OIF here is that the governance and safety infrastructure to deploy it responsibly are not yet in place.[8] However, we are only rejecting OIF in the near term and do endorse further research into it. The mixed results of trials between 1993 and 2009 are often seen as proof that OIF is not effective enough to matter, but most of those trials lasted only a few weeks, not long enough to determine if the carbon made it to the deep ocean before a bloom collapsed.[5] The only sufficiently long trial to observe real carbon export, EIFEX, suggests the limiting factor in earlier results was duration rather than mechanism.[6]

The 2026 modeling study poses a tougher problem: regardless of the trial length, OIF does not show uniformity across ocean regions.[10] For example, a version that works in the Southern Ocean could instead collapse zooplankton populations in the equatorial Pacific, making procedural justice difficult. One country might fertilize far-off waters, helping the climate, but another's fisheries pay the ecological price without an international body with the authority to govern that damage.[24]

There are also moral concerns to worry about. Resources that are spent on scaling a geoengineering fix are resources not spent on tackling emissions head-on. Additionally, deploying at a large scale could lead to termination shocks if the systems aren't maintained.[23] If future generations neglect these systems, they could be burdened with the

cleanup. None of this rules out OIF forever; it is simply calling for more modest and better-managed research before deployment.

4. Conclusion

Large-scale ocean iron fertilization is not an ethically justifiable solution to the climate crisis today. Amplifying the ocean's carbon pump offers a cheap theoretical way to pull carbon from the air. However, the unresolved risks of toxic algal blooms, deep-water oxygen loss, and downstream nutrient robbing make commercial or unilateral deployment far too dangerous.[10,15,16] Ethically, treating the global oceans as an open dumping ground for carbon mitigation violates the precautionary principle and invites moral hazard. It also threatens severe distributive injustice against nations whose fisheries and marine ecosystems could be wrecked without their consent.[22,23,24] Until strong international laws can guarantee procedural justice and bounded field trials can prove safe, lasting carbon capture, humanity must prioritize cutting emissions directly over geoengineering the seas.[8,14]

5. References

[1]Williamson, P., et al. "Ocean fertilization for geoengineering: a review of effectiveness, environmental impacts and emerging governance." *Process Safety and Environmental Protection*, 2012. https://doi.org/10.1016/j.psep.2012.10.007

[2]Martin, J.H. "Glacial-interglacial CO2 change: The Iron Hypothesis." *Paleoceanography*, 1990. https://doi.org/10.1029/PA005i001p00001

[3]Woods Hole Oceanographic Institution. "Iron Fertilization." https://www.whoi.edu/ocean-learning-hub/ocean-topics/climate-weather/ocean-based-climate-solutions/iron-fertilization/

[4]Smetacek, V., et al. "Deep carbon export from a Southern Ocean iron-fertilized diatom bloom." *Nature*, 2012. https://www.nature.com/articles/nature11229

[5]Yoon, J.E., et al. "Reviews and syntheses: Ocean iron fertilization experiments - past, present, and future." *Biogeosciences*, 2018. https://bg.copernicus.org/articles/15/5847/2018/

[6]Alfred Wegener Institute. "The European Iron Fertilization Experiment EIFEX." https://epic.awi.de/15162/

[7]Strong, A.L., et al. "Ocean fertilization: science, policy, and commerce." *Oceanography*, 2009. https://doi.org/10.5670/oceanog.2009.69

[8]International Maritime Organization (IMO). "Ocean Fertilization under the LC/LP." https://www.imo.org/en/OurWork/Environment/Pages/Ocean-Fertilization.aspx

[9]Tagliabue, A., et al. "The potential and risks of ocean iron fertilization." *Nature*, 2026. https://www.nature.com/articles/s41586-026-10795-y

[10]National Academies of Sciences, Engineering, and Medicine. "A Research Strategy for Ocean-based Carbon Dioxide Removal and Sequestration." 2022. https://nap.nationalacademies.org/catalog/26278/

[11]Blain, S., et al. "Effect of natural iron fertilization on carbon sequestration in the Southern Ocean." *Nature*, 2007. https://www.nature.com/articles/nature05700

[12]Sage Journals. "Perspectives on Ocean Geoengineering." *Sage Journals*, 2026. https://journals.sagepub.com/doi/pdf/10.1177/29768659261420631

[13]Assmy, P., et al. "Thick-shelled, grazer-protected diatoms decouple isotope fractionation from silicate depletion in a robust iron-induced bloom." *PNAS*, 2013. https://doi.org/10.1073/pnas.1309765110

[14]Buesseler, K., et al. "Scientists Make Case for Ocean Iron Fertilization Field Trials." *Dialogues on Climate Change*, Bigelow Laboratory, 2026. https://www.bigelow.org/news/articles/2026-02-05.html 7Srivastava, Rajiv Ranjan, and Sadia

[15]Law, C.S. "Predicting and monitoring the

effects of ocean iron fertilization on trace gases." *Marine Ecology Progress Series*, 2008. https://doi.org/10.3354/meps07545

[16]Trick, C.G., et al. "Iron enrichment stimulates toxic diatom production in high-nutrient, low-chlorophyll areas." PNAS, 2010. https://doi.org/10.1073/pnas.0910579107 Southern Ocean." *Science*, 2004. https://www.science.org/doi/10.1126/science.1086895

[17]Yu, J., et al. "Modeling the ecological impacts of marine carbon dioxide removal." *Earth System Dynamics*, 2026. https://esd.copernicus.org/articles/2026/

[18]Martin, J.H., et al. "Testing the iron hypothesis in ecosystems of the equatorial Pacific Ocean." *Nature*, 1994. https://www.nature.com/articles/371123a0

[19]Weier, J. "John Martin: Give me a half tanker of iron, and I will give you an ice age." NASA Earth Observatory, 2001. https://earthobservatory.nasa.gov/features/Martin

[20]Buesseler, K.O., et al. "The effects of iron fertilization on carbon sequestration in the

[21]Petit, J.R., et al. "Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica." *Nature*, 1999. https://www.nature.com/articles/20859

[22]Jamieson, D. "The Ethics of Geoengineering." *Environmental Values*, 1996. https://doi.org/10.3197/096327196776679159

[23]Lin, A.C. "Does Geoengineering Present a Moral Hazard?" *Ecology Law Quarterly*, 2013. https://escholarship.org/uc/item/7th0d0pd

[24]Preston, C.J. "Ethics and geoengineering: reviewing the moral issues raised by solar radiation management and carbon dioxide removal." *WIREs Climate Change*, 2013. https://doi.org/10.1002/wcc.204

Journal of Ethics in Scientific and Technological Innovations

The Ethics of Engineering for Performance versus Sustainability in Sports Equipment

Grace Zeng[1*]
[1]*International School, Bellevue, WA 98005*

Abstract

The field of sports equipment is constantly advancing to maximize the performance of athletes, especially that of professionals. However, optimizing for performance often results in environmental costs, especially at the end of a product's life. For example, the use of composite materials boosts results but complicates the recycling process. Currently, there are no standardized guidelines or regulations on the clash between performance and sustainability that are widely accepted.

This paper explores two examples encompassed within this topic, carbon-fiber composite tennis rackets and expanded polystyrene foam helmets, and utilizes them to inform a framework proposal to handle the competing interests of performance and sustainability. The nuanced framework it establishes considers several factors, including the magnitude and vitality of the performance gain offered compared to the severity of the environmental cost, and is a move toward recognizing and incorporating the variety of cases that pertain to this debate.

1. Introduction

Modern sports equipment is designed to optimize the performance of the athletes utilizing it. To achieve demanding performance requirements, engineers select and combine materials to produce specific properties such as strength, stiffness, energy absorption, and low weight.[1,2] Composite materials are one approach to achieving these properties and are widely used in high-performance sports equipment,[1,2] including tennis rackets, bicycle frames, helmets, and skis.

However, optimizing materials for performance can create challenges at the end of a product's life. Products made from multiple materials can be difficult to disassemble and recycle[1,2], while materials designed to undergo irreversible deformation may be difficult or impossible to reuse.[3,4] When materials cannot be recovered, their disposal contributes to waste and creates continued demand for virgin resources.[1,2] Thus, the properties that make a material effective during use can often conflict with its sustainability after use.

This issue is particularly nuanced because not all performance benefits carry the same significance. In a bicycle helmet, for example, the ability of a material to deform and absorb energy is directly responsible for protecting the wearer.[3] However, that same deformation can render the helmet unusable after a single impact[3] and is achieved with foam that poses a threat to the environment at the end of its life.[4] Meanwhile, in a tennis racket, carbon fiber composites provide performance advantages but are not essential to safety.[1,2] These cases raise a broader ethical question: when there is an inverse relationship between performance and sustainability, how should materials engineers determine which properties to prioritize?

2. Performance-sustainability tradeoffs in sports materials

2.1. Case Study: Carbon-fiber composite tennis rackets

Carbon fiber composites are used in professional equipment across a variety of sports, making up bicycle frames, tennis rackets, rowing boat shells, skis and poles, and more.[2] They are so ubiquitous because of the properties they possess: their composition makes them very light for their strength and stiffness.[2]

Carbon fiber composites feature strands of carbon atoms that run throughout a plastic resin. The resin is light and binds everything together, while the embedded carbon fibers distribute any mechanical stress, increasing the material's tensile strength and stiffness.

Our case study is on carbon-fiber composite tennis rackets. Carbon fiber accounted for more than 78.2% of the tennis racket market's global revenue in 2023,[5] suggesting that these composites dominate the performance-oriented segment of the market. The most popular alternative is aluminum, whose properties are less optimized for the sport, but is a lot cheaper.[1] This makes it a suitable material for beginner and recreational rackets.

To be exact, carbon fiber in an epoxy matrix has a lower density than aluminum at 1.5 g/cm3 compared to 2.8 g/cm3.[1] At the same time, it possesses a higher tensile strength of 2070 MPa compared to aluminum's 620 MPa and a higher Young's modulus of 138 GPa compared to 69 GPa.[1]

Carbon fiber's environmental cost stems from its energy-intensive production process and inherent recycling challenges.[1] In contrast to aluminum, which can be recycled infinitely in an energy-efficient manner without a reduction in quality,[1] carbon fiber composites are both expensive to recycle and experience a significant drop in quality that renders the product unsuitable for high-performance equipment.[1,2] Thus, over 95% of carbon fiber ends up in landfills.[6]

The intensive production process for the carbon fiber composite and its low recycling rate contribute to pollution and climate change and add to the global waste crisis, which harms humans globally.[1] These impacts occur within a broader global waste-management crisis that disproportionately affects lower-income countries. International trade in plastic waste often shifts difficult-to-manage waste toward lower-income countries that are unable to handle it in an environmentally sound manner, which can lead to pollution.[7]

However, it is important to note that tennis rackets only make up a small portion of the carbon-fiber industry as a whole. For reference, the American Composites Manufacturers Association estimated that sporting goods comprised 12% of the market use in 2020.[8]

2.2. Case Study: EPS foam bicycle helmets

Around 58% of commercial bicycle helmets rely on expanded polystyrene (EPS) foam.[9] Helmets protect the wearer's head in two ways: by slowing the impact so that the head stops over a longer period of time, thus reducing the force the head experiences, and by converting a portion of the energy from the collision into heat through deformation.[3] The EPS foam in the helmet both slows the crash down and deflects energy from the impact when many of the tiny closed cells within the foam rupture.[3]

Another type of foam that is seen in some helmets is expanded polypropylene (EPP). EPP foam is able to provide impact protection over multiple collisions, recovering its shape well.[3] However, EPP foam helmets are mainly produced for multi-impact sports such as skateboarding, and not as much for bicycling.[3] This is because EPS is simply more suitable for ensuring safety in severe crashes.[3] Similarly, other rubbery multi-impact foams such as thermoplastic polyurethane (TPU) are also

largely restricted to other sports, also partially because of their higher weight.[3]

Currently, 58% of bicycle helmets rely on EPS foam[9] because it is good at ensuring safety in severe impacts in addition to being light and cheap.[3,10] However, EPS foam poses a severe threat to the environment.[3,4,10]

A single impact that crushes the cell walls severely impairs the ability for the helmet to continue ensuring safety in another crash.[3] Thus, consumers are required to discard and replace the helmet after every crash.[3]

Moreover, it is difficult to properly dispose of EPS helmets at the end of their life in a manner that does not cause harm to the environment.[4,10] Expanded polystyrene does not biodegrade in the environment; rather, it is only able to break up into increasingly small pieces.[4] Over time, a piece of EPS foam transforms into microplastics that are harmful to animals, who are unable to digest them.[4] This harmful effect is exacerbated by EPS foam's low density, which allows it to easily get picked up and travel long distances with the wind.[4,10]

Consumer EPS recycling programs are rare because EPS foam is difficult to recycle and the process is more expensive than simply creating new, higher-quality EPS foam from virgin material.[4,10] Recycling it is complicated because of the spectrum of chemicals that are added to modify its properties and appearance[4] and how easily it gets contaminated.[10] EPS foam is also highly inefficient to transport due to its low density.[4]

The design of bicycle helmets as a whole is also non-conducive to their sustainability. Helmets are composed of layers of different materials for the outer shell, shock-absorbing liner, and other components. These layers are traditionally fused together when manufactured using glue or other techniques, which helps them stay together during crashes. However, it also makes it challenging for users and recycling facilities to separate and recover any of the materials.[11] Thus, after a crash, the entire helmet must be replaced due to safety concerns around the EPS foam liner, and in the vast majority of cases, the whole helmet is discarded in a landfill or incinerated.[11]

3. Ethical frameworks

3.1. Current engineering ethical codes

Many professional engineering organizations such as the National Society of Professional Engineers (NSPE), the American Society of Mechanical Engineers (ASME), and the American Institute of Chemical Engineers (AIChE) mandate that their engineers follow their own code of ethics.[12,13,14] These serve as guidelines within which all professional work must be performed.

These codes all emphasize that engineers must "hold paramount the safety, health, and welfare of the public".[12,13,14] They also all promote sustainability, though to different extents. The NSPE encourages them to "adhere to the principles of sustainable development" and considers it an extension of serving the public interest.[12] The ASME dictates that "engineers shall consider environmental impact and sustainable development".[13] The AIChE uses more binding language by mandating that their engineers "protect the environment".[14]

Overall, the engineering codes of ethics in use today tend to imply through the wording and order of edicts that ensuring safety is paramount while protecting the environment is important, though not as essential.[12,13]

3.2. Relevant ethical theories

3.2.1. Consequentialism

Consequentialism urges people to make decisions based on the consequences that would ensue from a choice.[15] It dictates that decision makers should choose the option that would lead to the best consequences overall.[15]

Thus, consequentialism would urge sports equipment engineers to consider any and all relevant factors including, but not limited to, the user's safety, the product's relative performance advantages or disadvantages, the environmental impact the equipment has, and the product's cost. The ethically preferable choice would be the one that produces the greatest overall benefit or least overall harm.

3.2.2. Deontology

Deontology is founded upon the idea that there are duties, or moral obligations, that people must follow.[15] It differs from consequentialism in that it does not consider the overall consequences of the action, only whether the action itself is right or wrong.[15]

According to deontological principles, product designers may have a duty to protect users from preventable harm. In the bicycle helmet case study, this duty could justify prioritizing a material's protective performance over sustainability when a more sustainable alternative would provide inadequate protection.

3.2.3. Distributive justice

Distributive justice evaluates the morality of decisions based on how fairly the benefits and the costs associated with them are distributed.[16]

In the context of sports equipment, engineers should consider who benefits from a material's performance advantages and who bears its environmental costs. For example, a high-performance material provides performance benefits to athletes who can afford it, while environmental costs associated with its production and disposal may be distributed across society more broadly.

4. Proposed framework

This paper recommends that, when evaluating which factors to optimize for in sports equipment, engineers recognize the existence of a spectrum of how essential improved performance is.

More specifically, the ethical acceptability of an environmentally costly material should depend on the following three factors.

1. if it is able to meet baseline safety and functional standards while less environmentally costly alternatives are not
2. the magnitude of the performance advantage it offers, weighed against the additional environmental harm it causes
3. the distribution of the resulting performance benefits and environmental costs

To begin, a material choice should not sacrifice adequate safety or essential performance for any environmental benefit. Adequate safety is defined based on the established industry safety standards for a given product. Essential performance is defined as the performance necessary for a product to fulfill its fundamental function safely and effectively, as opposed to performance that provides only an incremental competitive or consumer advantage.

Should multiple materials satisfy these minimum requirements, the next step is to weigh the magnitude of the performance advantage offered by each material against the additional environmental harm it may bear. The greater the environmental cost of a material choice, the stronger its performance benefit has to be to justify the cost. Here, performance also includes additional safety benefits. These must be seriously considered because even if a piece of equipment meets the minimum safety threshold, manufacturers may face product-liability claims if an alternative design could have provided greater protection outweighing associated sacrifices and the chosen design is subsequently alleged to be unreasonably unsafe.[17]

The last factor to be considered is the relative distribution of the benefits and costs that

come with a material. Environmental costs are generally felt across the globe and by people of all socioeconomic statuses and backgrounds. If the equipment is accessible and widely used, meaning that the performance advantages are broadly distributed, it would be more fair than if it were restricted to a small group, because the environmental costs are widely experienced as well.

This framework incorporates the regard for safety and, to a lesser extent, the environment, from current ethical codes of professional engineering organizations. By weighing the various consequences that arise from decisions, it mirrors the holistic approach of consequentialism. It prioritizes the user's safety, which deontology might deem a duty. And it considers that environmental costs often pertain to a wider audience than performance gains, a concept of distributive justice.

4.1. Proposed framework when applied to carbon-fiber composite tennis rackets

Carbon-fiber composites offer a significant performance upgrade over aluminum due to their reduced weight and durability. However, the performance of tennis racket frames is not essential to the safety of their user. Aluminum rackets are also a passable alternative and are often employed by beginners and recreational users. Due to the lack of widespread and feasible recycling programs, carbon fiber has a much greater impact on the environment than aluminum.

Additionally, because carbon fiber rackets are premium and also come with a higher price tag, they are used by a small subset of people who play tennis and are relatively skilled. However, the environmental cost of the use of carbon fiber unfairly affects people who do not use the rackets.

It is important to acknowledge that the environmental impact of carbon-fiber tennis rackets is limited relative to the carbon-fiber industry as a whole. Even when considering all sporting equipment, this sector accounted for approximately 12% of global carbon-fiber use in 2020.[8] Therefore, reducing the use of carbon fiber in tennis rackets alone would have a relatively limited effect on the overall environmental impact of carbon-fiber production.

However, the framework proposed in this paper is intended to encourage a broader approach to material selection, rather than apply only to sports equipment. Applying this mindset across industries can help engineers consider whether the performance benefits of a material are sufficient to justify its environmental costs, particularly when lower-impact alternatives are available. Thus, while the environmental impact of an individual material choice may be limited in scale, carefully evaluating such choices remains important as part of broader efforts toward more sustainable engineering.

Considering all of the aforementioned points, it is difficult to justify the environmental cost of carbon fiber tennis rackets for the average recreational player. Aluminum rackets suffice for casual use in less skilled environments because recreational players do not require the best performing materials if they are playing simply for their health and enjoyment. However, exceptions may be considered for elite competitive players, because at such a high level, the magnitude of the performance advantage offered by carbon fiber becomes more significant.

4.2. Proposed framework when applied to EPS foam bicycle helmets

EPS foam is essential for the safety of bicycle riders in severe accidents. Its performance advantage is directly related to safety. Additionally, it does not have a competitor that simultaneously possesses similar enough properties and a lower environmental impact.

Finally, due to the widespread popularity of biking and the low cost and prevalence of EPS foam helmets, the safety advantages extend to a larger pool of people. This renders the environmental costs, which are also widely distributed, more fairly spread.

Therefore, while EPS bicycle helmets pose an environmental threat, it does not justify replacing them with an alternative material.

5. Conclusion

This paper determined that there is no universal answer to whether prioritizing performance or sustainability is ethically correct. Instead, a framework must be established to determine which should be optimized, taking into account the magnitude and vitality of the performance increase compared to the severity of the environmental cost, evaluated through a comparison with existing alternative options.

These results are significant because proponents of either perspective, particularly sustainability, may sometimes argue that their preferred factor should be prioritized across all circumstances. In doing so, they can oversimplify complex tradeoffs by emphasizing examples that support their position while overlooking cases in which other factors might reasonably take priority.

Future work includes building upon this framework by providing more absolute guidelines, instead of relying on isolated case studies and metrics that may leave room for interpretation.

6. References

[1]Subic, Aleksandar, et al. "Sustainable Design and Environmental Impact of Materials in Sports Products." Sports Technology, vol. 2, no. 3–4, June 2010, pp. 67–79, https://doi.org/10.1002/jst.117.https://doi.org/10.1038/d41586-023-03974-8.

[2]Temesgen, Alhayat Getu, et al. "Advanced Composite Technologies for High‐Performance Sports Equipment: Materials, Manufacturing, and Sustainability Perspectives." Polymers for Advanced Technologies, vol. 37, no. 5, Wiley, May 2026, https://doi.org/10.1002/pat.70605.

[3]"Bicycle Helmet Liners: Foam and Other Materials – BHSI." Helmets.Org, 5 May 2026, https://helmets.org/bicycle-helmet-liners-foam-and-other-materials/. Accessed 23 Aug. 2026.

[4]Beyond Plastics. "Beyond Plastics - Working To End Single-Use Plastic Pollution." Beyond Plastics - Working To End Single-Use Plastic Pollution, 12 Feb. 2024, https://www.beyondplastics.org/fact-sheets/polystyrene. Accessed 24 Aug. 2026.

[5]Research and Markets ltd. "Research And Markets." Researchmarkets, 2024, https://www.researchandmarkets.com/report/racket?srsltid=AfmBOoqNWh-RVl16gUVn4FwNgyzXygulB6DtTvteQ-xANd7_pb8jjprl. Accessed 22 Aug. 2026.

[6]Bruniaux, Anastasia. "Closing the Loop on Carbon Fibre Waste - Trailblazer for Recycling & Clean Energy." Trailblazer for Recycling & Clean Energy, 20 Apr. 2026, https://trace.org.au/closing-the-loop-on-carbon-fibre-waste/. Accessed 26 Aug. 2026.

[7]"The OECD Has Been Monitoring Trade in Plastic Waste and Scrap since 2021 and Publishes Annual Monitoring Reports. OECD's Work on Plastics Responds to Growing Global Concerns about the Environmental and Human Health Impacts of Plastic Pollution.." OECD, 22 July 2025, https://www.oecd.org/en/data/insights/data-explainers/2025/07/why-is-the-oecd-monitoring-the-trade-of-plastic-waste.html. Accessed 9 Sept. 2026.

[8]"2021 State of the Industry Report." American Composites Manufacturers Association, 15 Feb. 2021, https://acmanet.org/2021-state-of-the-industry-report/. Accessed 23 Sept. 2026.

[9]"Bike Helmet Market Size | 2024 To 2032 Analysis Report."

Businessresearchinsights.Com, 2024, https://www.businessresearchinsights.com/market-reports/bike-helmet-market-102219. Accessed 25 Aug. 2026.

[10]"Can Polystyrene Be Cost-Effectively Recycled? - iSustain." Isustainrecycling.Com, 2015, https://isustainrecycling.com/blog/post/can-polystyrene-cost-effectively-recycled. Accessed 26 Aug. 2026.

[11]"Recycling Bicycle Helmets and Sustainable Design – BHSI." Helmets.Org, 12 May 2026, https://helmets.org/recycling-bicycle-helmets-and-sustainable-design/. Accessed 26 Aug. 2026.

[12]NSPE. "NSPE Code of Ethics for Engineers | National Society of Professional Engineers." Nspe.Org, National Society of Professional Engineers, 2019, https://www.nspe.org/career-growth/nspe-code-ethics-engineers. Accessed 25 Aug. 2026.

[13]ASME. "Code of Ethics of Engineers." Www.Asme.Org, American Society of Mechanical Engineers, Oct. 2024, https://www.asme.org/getmedia/1d9d64bf-f6a7-4e75-a8f5-0cabd8c04269/p-15-7-ethics_2.pdf. Accessed 25 Aug. 2026.

[14]AIChE. "Code of Ethics." Www.Aiche.Org, American Institute of Chemical Engineers, Nov. 2015, https://www.aiche.org/about/governance/policies/code-ethics. Accessed 25 Aug. 2026.

[15]Kühler, Michael, et al. "Ethical Theories." The Routledge International Handbook of Engineering Ethics Education, Routledge, Jan. 2025, pp. 44–59, https://doi.org/10.4324/9781003464259-4. Accessed 24 Aug. 2026.

[16]Kaufman, A. "Distributive Justice - an Overview." Sciencedirect.Com, 2017, https://www.sciencedirect.com/topics/social-sciences/distributive-justice. Accessed 24 Aug. 2026.

[17]Cornell Law School. "Product Liability." LII / Legal Information Institute, 2024, https://www.law.cornell.edu/wex/product_liability. Accessed 23 Sept. 2026.

Journal of Ethics in Scientific and Technological Innovations

A Review of the Mechanics, Efficacy, and Ethical Concerns Associated with Enhanced Rock Weathering

Xuanming Liu[1]
[1]*Syosset High School, Syosset, NY 11791*

Abstract

Enhanced rock weathering (ERW) is a promising form of carbon capture that involves the mining, grinding, and scattering of ground silicate rocks such as basalt or ultramafic rocks. These rocks, containing magnesium or calcium oxide, can react with atmospheric carbon dioxide in order to form bicarbonate and eventually calcium carbonate in the form of calcite or magnesium carbonate in the form of magnesite. This method can absorb carbon directly from the atmosphere or seawater and store it in a stable form for thousands of years. It also yields numerous potential co-benefits including, enhanced crop yields, improved soil quality and hydrology, and alleviating ocean acidification. However, the processing and mechanics of this method have numerous ethical and scientific concerns that must be addressed before full scale implementation. For example, ultramafic rocks, and iron slag can contain heavy metals that may accumulate in soils and potentially get absorbed by crops. So mitigation strategies like hyperaccumulators should be considered to target potential health and food security issues. Other issues involve air pollution such as PM2.5 silica dust from grinding, which may cause both environmental damage and harm to the workers especially in poorer countries where less safety structures are implemented. Less wealthy nations could also be initially subject to environmental inequality due to the large initial costs and power requirements of the technology; however models have shown that many nations such as India and Brazil can overtake western nations with some preemptive support. Ecosystem damage through formation of secondary byproducts or leached pollutants along with the discouragement of minimizing emissions are also potential challenges.

1. Introduction

Climate change is a continuously worsening issue, and global greenhouse gas emissions must be lowered to combat issues facing all aspects of life. Net zero CO_2 energy systems entail both a reduction in fossil fuel use and the use of carbon capture technologies to offset remaining fossil fuels.[1] Enhanced rock weathering (ERW) is a promising carbon capture method that speeds up a natural atmospheric CO_2 trapping process. Atmospheric CO_2 can be absorbed by rainwater, forming carbonic acid. This carbonic acid slowly weathers silicate rocks and forms stable bicarbonate, effectively storing and disposing CO_2. Naturally occurring silicate weathering reactions are estimated to bind 100 million tonnes of carbon every year.[2] ERW involves the mining, grinding, and spreading of these silicate rocks to enhance their weathering rate by increasing surface area. Basalt and ultramafic rocks such as dunite are typically considered suitable for ERW, since they are abundant in minerals such as olivine which readily dissolves in CO_2 rich water to form bicarbonate and can eventually form solid magnesium carbonate.[3, 4]

Global modeling studies have predicted that Dunite has the potential to remove up to 95 $GtCO_2$ every year and basalt can remove up to 4.9 $GtCO_2$ every year, and in warm humid areas up to 75% of the global potential can be reached.[5] Through the use of predictive models based on local wealth, infrastructure, resources, and social tipping points, by 2028 the

sequestration rate of CO_2 is projected to stabilize between 0.7 and 1.1 Gt each year collectively removing 34-55 $GtCO_2$ by 2100.[6]

Enhanced rock weathering is also a promising solution due to its many co-benefits. For example, it can lessen ocean acidification issues, stabilize soil pH, provide crucial nutrients to increase plant yield, and improve soil hydrology.[7, 8, 9, 10]

However, the method presents many drawbacks including atmospheric pollution caused by the mining and grinding of silicate rocks to scatter on farmland and in the oceans. The fine silica dust could be a hazard to workers especially in less wealthy nations which have less safety regulations and infrastructure. Heavy metals such as Nickel and Chromium could be incorporated into the vegetables grown on farmland used for ERW. Global scaling could be difficult as well, since poorer nations may lag behind western developed nations leading to environmental inequality. Seemingly having these carbon capture methods could also be seen as a safety net and discourage alarm on fossil fuel usage. ERW by itself will not be able to significantly reduce net CO_2 output if emissions are not effectively minimized.[6, 11, 12]

2. Methods of Enhanced Weathering

Enhanced rock weathering can be performed using a variety of different methods. For example, ground basaltic and ultramafic rocks containing olivine can be spread over agricultural fields or forests to increase yield and provide essential nutrients after being broken down.[9, 16] Olivine can be also dissolved in the oceans in order to sequester CO_2, while increasing the alkalinity of the oceans by consuming protons. A stoichiometric maximum of 1.25g of CO_2 could be absorbed by the oceans for every 1 gram of olivine dissolved,[7, 13, 14] although dissolution rate remains a concern.[15] Industrial byproducts and waste such as iron slag or cement can even be used for ERW.[17, 18] Figure 1 summarizes the terrestrial and marine transportation pathways after the rock dust is spread.



Figure 1. Graphical overview of terrestrial and marine enhanced rock weathering transportation pathways. Reproduced from Levy et al., Environmental Science & Technology, 2024, 58, 17215–17226, https://doi.org/10.1021/acs.est.4c02368, under a CC-BY-NC-ND 4.0 license (https://creativecommons.org/licenses/by-nc-nd/4.0/).[54]

2.1 Sourcing and extraction of rock powder

Enhanced rock weathering does not need new technologies like other carbon dioxide removal methods; instead, ERW can employ existing agricultural machinery to spread the rock powder and currently tap unused stocks of mined and quarried rock. These rocks are then ground and transported to locations where ERW will be implemented. However, as it continues to expand, waste stocks will run out and many new rocks will need to be newly quarried and mined causing environmental damage.[34] Economic, environmental, and social sustainability could vary widely, and environmental concerns could be major in countries with weaker regulatory laws.[35]

2.2 Scattering of Silicates Terrestrially

Terrestrial scattering involves basalt or ultramafic rock powder being spread by standard agricultural spreading equipment such as agricultural lime spreaders.[33] Climate model simulations predict that if 10 tons of basalt dust per hectare was scattered on only 1000

agricultural sites, identified with a 1D reactive transport model, globally, 64 gigatons of CO_2 would be absorbed over 75 years, and if ERW was applied to all agricultural land 217 gigatons of CO_2 would be absorbed.[19] Terrestrial scattering of silicates for ERW is especially effective in tropical latitudes and India due to their strong incoming solar radiation and precipitation from the Intertropical Convergence Zone and Indian Monsoon, since rainwater is the primary solvent for the dust and temperature speeds up reactions.[19] This trend occurs in forests as well, although biotic ERW accounts for 23.4% of the total carbon capture occurring in forests, and 66.0 $MtCO_2\ yr^{-1}$ is absorbed across simulated US managed forests over 10 years.[16] Spreading the rock powder in forests would be more difficult, however; vehicles can only spread the powder where they can drive, and aerial spreading could lead to waste due to dense canopies. The model created in this study defined fine-grained basalt as particles with an initial p80 value, the diameter where 80% of the total basalt mass is made up of particles with a smaller diameter, of 100 micrometers, while coarse-grained basalt had a p80 value of 1220 micrometers. Finer grains of basalt greatly improved the effectiveness of ERW since larger grains in agricultural fields can oftentimes be unweathered even after 75 years.[19]

2.3 Scattering of Silicates in Marine or Coastal Environments

The mechanisms of marine olivine weathering and its effect on seawater is still poorly understood and more monitoring, reporting, and verification (MRV) is needed. Finely ground olivine is scattered onto coastal and shelf environments, where it can be integrated into necessary beach nourishment or land reclamation due to coastal erosion and rising sea levels.[24] ERW could be applied to the open ocean as well, but it would require much finer grinding of the basalt, requiring more energy usage.[13] Experiments have shown that the dissolution rate of olivine in seawater is uncertain, and it dissolves nonstoichiometrically making prediction difficult.[15] Dissolution in filtered seawater also deviates significantly from artificial seawater, so more real world testing is required before implementation.

Project Vesta sought permits for a pilot placing of olivine sand at 2% of the volume of a beach nourishment project off Duck, North Carolina and 5% at Southampton NY. So far after two years of monitoring, no significant ecological impact has been found and no significant impact on key marine organisms such as amphipods and polychaete worms was found in toxicity tests. However this is only one test that is still very early in its lifetime, and more information is needed before deployment of carbon capture in the oceans.[30, 31]

2.4 Use of Industrial Byproducts

Cement waste and iron slag are two promising materials for enhanced rock weathering due to their faster degradation times, while also utilizing industrial waste that would otherwise go unused. However, the effectiveness of these byproducts are limited by variable CaO content for cement waste and heavy metal contributions such as Cr, Cu, and Pb.[17, 18, 25] Laboratory experiments have shown that cement waste was able to effectively absorb CO_2 and directly form calcite, changing its composition from 7% to 18%; however, removing 1 tonne of CO_2 requires 20 tonnes of cement waste.[17] Cement waste can be mixed with other materials such as basalt to provide additional increases in the alkalinity of soil.[17] Iron slag had no damage to organic soil matter including minerals associated and particulate along with not impacting surface CO_2 fluxes.[18] Estimates suggest that iron and steel slag have global CO_2 reduction potentials of 138 million tons per year, since over 400 million tonnes of iron and steel slags are produced each year.[32] 7-17 billion

tonnes of cement waste are produced globally each year having the potential to sequester 190-332 million tonnes of CO_2.[33] Both cement waste and iron slag still need more MRV as well before they can be fully implemented as their side effects and effectiveness are still not completely understood due to a lack of testing.

3. Co-benefits of Enhanced Weathering

3.1 Alleviates Ocean Acidification

Ocean acidification is extremely harmful to ocean ecosystems, since it can thin shells made out of calcium carbonate, stunt growth, and destroy coral reefs which harms an entire ecosystem.[23] Ocean acidification is due to overabsorption of CO_2 creating carbonic acid which then forms bicarbonate and hydrogen ions leading to a 30% increase in ocean acidity.[26] The dissolution of olivine in seawater can absorb hydrogen ions created in the reaction to form bicarbonate in the oceans. olivine dissolution acts as a buffer and increases the alkalinity of seawater.[15]

Terrestrial scattering of basalt or ultramafic rock powder's basic bicarbonate and carbonate molecule formation from atmospheric CO_2 can flow into the ocean through rivers, rainwater, and groundwater transport. This influx of basic molecules can neutralize some of the hydrogen ions in seawater alleviating ocean acidification.[7]

3.2 Stabilizes Soil pH

Similar to the effects enhanced rock weathering has on the pH of the oceans, ERW can also increase the soil pH in fields where the rock dust is scattered. The absorption of CO_2 by basalt or ultramafic rocks also absorbs the hydrogen ions in the soil, and in past experimentation soil pH increased from 6.2 to 6.8 between 2017 and 2020 for maize at 0–10 cm depth, 5.6–6.0 in miscanthus.[18] Many basic cations such as magnesium or calcium can be released as well. The increase in soil pH can also form amorphous Al(OH)3, SiO2, or Fe(OH)3 secondary mineral phases which reduces the transport of basic ions away from the fields utilizing enhanced rock weathering.[27] From a three year experiment in the US Midwest Iron slag proved evidence of removal of atmospheric CO_2 in acidic soils and can increase their pH to safer levels for plants while traditional ERW materials performed worse; however, in neutral pH levels iron slag had no significant results.[18]

3.3 Provides nutrients and increases yield

Real world tests of ERW on farmland have achieved 21±9.4% of increased yield of *Sorghum bicolor*, one of the most cultivated crops globally, due to basalt containing potassium, phosphorus, calcium, magnesium, iron, and manganese which are all plant-essential nutrients and low concentrations of toxic nickel and chromium.[13, 20, 21] Basalt also contains silicon which plants accumulate in their straw and stover to protect against external stresses, and experiments have shown that silicon concentrations increased by 26±5.4%.[13, 22] Field experiments have shown that ERW also offsets carbon loss in maize/soybean by 23%-42% and increased inorganic carbon storage of miscanthus which had an initial storage rate of 62-129$gCm^{-2}yr^{-1}$ by 234$gCm^{-2}yr^{-1}$.[8] This could decrease the reliance on toxic fertilizers for increasing yield and provide an environmentally friendly option instead.

3.4 Improves hydrology

Silicate dust from ERW can improve rainwater retention, optimizes pore size distribution, and reduces surface runoff. Experimentation has shown that enhanced rock weathering through rock powder can increase plant-available water through the larger pores of the basalt powder up to 31% from 13.8% to 18.2%.[21, 10] This practice could potentially help

mitigate the effects of droughts on crops, which are becoming more and more important as temperatures continuously rise due to carbon emissions. By improving topsoil water retention, enhanced weathering can alleviate drought impacts in regions used for afforestation, reforestation, and bioenergy grass plantations as well.[28] As the rock powder weathers clay minerals can be formed which then potentially increases the overall water holding capacity even further.[29] However, the impact depends heavily on the texture of the basalt and initial soil, where fine powder improves plant-available water while a coarse powder could even decrease it in certain areas.[10]

4. Environmental and Safety Risks

4.1 Air Pollution and silica dust inhalation

Currently there are no epidemiological studies of active ERW workers, so the evidence below comes from related settings and only indicates possible risks. Rock dust from enhanced rock weathering may be transported through erosion spreading rock particles to nearby areas and farmland, affecting air quality. Natural basaltic dust in Iceland has been found to reduce air quality and increase hospital admissions.[36] Studies of mineral dust suggest exposure burdens may be greater where controls are weaker[37]. Exposure to carbonate mineral dust can also lead to pneumoconiosis and lung inflammation in farmers and workers.[38, 39] Whether ERW silicate powders will have similar results is still unknown. The risk of lung cancer was elevated in workers that deal with silica dust with a determined standardized mortality ratio of 2.32 for silicotics and 1.78 for nonsilicotics with miners being at the highest risk.[42] However, by analogy, workers grinding mined rocks into fine dust may face comparable or elevated risks. Miners, farmers, and drivers working on collecting and using basalt or ultramafic rock could face respiratory risks if their exposures resemble those in related industries.

This could lead to increased risks, especially in lower income nations with far less safety structures in place if ERW becomes a global carbon capture solution. The specific air pollution and dust inhalation concerns have not yet been extensively studied, and most health impacts come from adjacent topics. More MRV is needed as well to understand the true potential danger of ERW, but health and ethical concerns will need to be addressed specifically for ERW due to its unique form of widescale spreading.

Enclosed or wet-milling grinding systems with extensive ventilation and filtration or integrated dust suppression systems with ultrasonic atomization with acoustic agglomeration could be plausibly adapted to ERW grinding facilities.[43] Slurry application of rock powder could decrease airborne particulates greatly but would also be more challenging to spread globally due to decreased usage of existing technology. PPE, remote operation, and exposure limits guidelines are likely needed and should be written into carbon-credit certification and MRV protocols to ensure equality even in less fortunate areas. However, this could lead to less implementation in less developed nations leading to environmental injustice. Epidemiological studies into ERW workers could be extremely beneficial and provide crucial information about the feasibility and issues with global scaling of this technology.

4.2 Heavy metals in plants

Simulations have suggested that if basalt was applied at 40 tonnes per hectare, regulatory limits could potentially be exceeded in as low as 6 years for copper and 10 years for nickel. However, other experiments have measured that even extremely high levels of basalt application such as 100t ha^{-1} did not lead to elevated contents of nickel copper or chromium in Sorghum bicolor.[9] All soils have some amount of toxic trace elements, which can have their concentration increased when powdered rock

undergoes rapid chemical weathering while other components endure in soil for decades. These results are not directly contradictory. Soil accumulation[44] does not necessarily mean plant uptake,[9] and the higher soil pH produced by ERW may reduce metal bioavailability.[45] Sourcing, application rate, and weathering rate all impact the toxic element accumulation from basaltic rock powder.[44]

Copper can accumulate strongly in upper layers of the soil and doesn't leach as much as other metals. In other agricultural contexts, copper accumulation has been linked to reduced productivity and food-safety concerns.[45, 46] ERW deployment must be optimized with careful selection of sites to collect material with less toxic metals and ensure relevant soil protection policies. Naturally formed basaltic and ultramafic rocks should be studied for more information, and the mobility or bioavailability could be limited by the higher pH levels created by ERW.[45] Hyperaccumulator crops can also be put into a crop rotation to ensure that toxic metals are extracted from the soil since they can reach concentrations of 1-3% of toxic metals due to their biological adaptation to prevent being consumed.[47]

Heavy metal risk is therefore best treated as conditional and manageable: it can be reduced by selecting low-metal feedstock, setting application limits, and monitoring soils and plant tissue, but long-term field data are needed to confirm this.

4.3 Environmental Ecosystem Harm

In general, deposition of particles less than ten micrometers can coat plants leading to less photosynthesis and eventually causing death. Lichens, liverworts, and sphagnum can be vulnerable as well leading to ecosystem damage.[40] These environmental pollutants can alter nutrient cycling and availability along with affecting microbes that live in acidic soils or environments such as swamps and bogs.[41] Freshwater environments can also be sensitive to pH changes. Mining can also lead to habitat destruction, and the mining, grinding, and transport of rock dust could lead to produced emissions. Overall, many conducted studies report positive or neutral effects such as phytoplankton reactions to olivine dissolution, although heavy metal pollution could be a potential large issue, or soil quality.[48, 49, 50]

5. Environmental Injustice and Control

This is also a fairly new topic for ERW, and doesn't have many studies but could be of great importance for global scaling. The regions near quarries and processing facilities that are directly impacted by the issues of the technology could often benefit less from the carbon credits, instead the benefits go to the companies and shareholders running the ERW operations and not the people put in potential dangers.[51]

A study computationally modeling the adoption of ERW shows that lower-income regions such as the Middle East, North Africa, South Asia, and Sub-Saharan Africa would likely face adoption delays of 13 to 15 years due to less existing infrastructure, guidelines, and funding.[6] However, some countries in these regions are projected to overtake Western nations over time.[6] By 2100 low and lower-middle income countries are projected to rise to 60% of the carbon removal.[6] If technology, knowledge, and funding structures work as intended, there would be equitable outcomes and spread benefits around the world. However, this would only occur if precautions are set in place. If deployment is profit-driven even amidst ongoing climate change, the necessity of rock powder processing, transportation, and scattering could prevent nations from contributing and benefiting. And as this article shows the effectiveness of ERW depends on these lower income agricultural areas near the equator. Policy must be ambitious and properly deployed globally and access to

funding, and knowledge is crucial. Western nations that adopt first must help other nations achieve the same in order to create net zero emissions.

Post-implementation is an issue as well for the ethical debate. The responsibility for continuation and monitoring will need strong structures to maintain.[52] This could be costly, and there could be risks of an abrupt halt to geoengineering efforts.[52] ERW is presented as a nonstop method and the effects might not last if it is abruptly stopped and the side effects would be completely unknown if maintenance was suddenly paused.[12] This is a large issue that needs to be planned ahead of implementation to ensure smooth control.[12]

Collective action risks failure if states pursue their own interests, since some states could take no action and benefit from an improved world, creating incentives to be an inactive state.[12] Especially if risk is placed onto other nations; for example, if hot humid climates are shifted to be the focus of ERW, underprivileged people in those nations would be the ones exposed to the risk.[12] Countries in drier, colder climates would gain the benefits but not have to contribute to implementing these solutions.[53] By the moral hazard, countries that feel like they have the safeguard of removing carbon with ERW may diminish efforts to reduce greenhouse gas emissions if thought of as insurance.[12] And realistically from carbon absorption rates, ERW will need to be paired with lowering carbon emissions.[6] If ERW is perceived as a safety net, it could potentially create less urgency to cut fossil fuel use.

6. Conclusion

ERW is a credible but unproven supplement to emission cuts, and it should be scaled gradually under strict monitoring. Enhanced rock weathering offers a high potential of carbon absorption from prebuilt technology combined with many possible co-benefits such as improved crop yield, soil pH, water retention, and reduced ocean acidification that most CDR approaches do not offer. It produces stable products in the form of standard rocks and minerals that trap carbon for thousands of years without continual interference. Models suggest that 34-55 $GtCO_2$ can be removed by 2100.[6] Large agronomic gains in plants such as Sorghum bicolor yield,[13, 20, 21] and a promising method for lessening ocean acidification[7, 15] are all possible benefits of ERW. ERW can help the industries where it is implemented.

However, ERW cannot be separated from its unresolved risks that still require research, real world testing, and understanding before this solution can be implemented. Epidemiological studies,[42, 43] heavy metal accumulation,[44, 47] and ecosystem effects[30, 31] are still in very early research stages.

Many ethical and structural concerns stand as well such as environmental injustice towards tropical, humid, lower-income regions.[6, 19] The populations most exposed to the potential dangers are not the ones gaining economic upsides in the form of carbon credits.[51] Lower income nations are necessary for effective implementation, contributing 60% by 2100,[6] so investments into infrastructure, safety, and regulation are also necessary before scaling globally. ERW will only be one part of the global climate solution. It can lead to large changes but only if done correctly, only if knowledge, technology, and funds are shared and only if emissions are continuously reduced.[6, 12] Further targeted research and planning are necessary, but if done correctly, with proper MRV, regulation, and safety measures, this technology could be beneficial to net zero emissions and slowing the negative effects of climate change.

7. References

[1] IPCC. (2023). Summary for policymakers. In H. Lee & J. Romero (Eds.), Climate change 2023: Synthesis report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (pp. 1–34). IPCC. https://doi.org/10.59327/IPCC/AR6-9789291691647.001

[2] SEIFRITZ, W. "CO2 Disposal by Means of Silicates." Nature, vol. 345, no. 6275, June 1990, pp. 486–86, https://doi.org/10.1038/345486b0.

[3] Schuiling, R. D., and P. Krijgsman. "Enhanced Weathering: An Effective and Cheap Tool to Sequester Co2." Climatic Change, vol. 74, nos. 1–3, Jan. 2006, pp. 349–54, https://doi.org/10.1007/s10584-005-3485-y.

[4] Vienne, Arthur, et al. "Enhanced Weathering Using Basalt Rock Powder: Carbon Sequestration, Co-Benefits and Risks in a Mesocosm Study With Solanum Tuberosum." Frontiers in Climate, vol. 4, May 2022, https://doi.org/10.3389/fclim.2022.869456.

[5] Strefler, Jessica, et al. "Potential and Costs of Carbon Dioxide Removal by Enhanced Weathering of Rocks." Environmental Research Letters, vol. 13, no. 3, Mar. 2018, p. 034010, https://doi.org/10.1088/1748-9326/aaa9c4.

[6] Tu, Ying, et al. "Scaling up Enhanced Rock Weathering for Equitable Climate Change Mitigation." Communications Sustainability, vol. 1, no. 1, Feb. 2026, https://doi.org/10.1038/s44458-026-00034-w.

[7] Taylor, Lyla L., et al. "Enhanced Weathering Strategies for Stabilizing Climate and Averting Ocean Acidification." Nature Climate Change, vol. 6, no. 4, Dec. 2015, pp. 402–06, https://doi.org/10.1038/nclimate2882.

[8] Kantola, I. B., et al. "Improved Net Carbon Budgets in the US Midwest Through Direct Measured Impacts of Enhanced Weathering." Global Change Biology, vol. 29, no. 24, Aug. 2023, https://doi.org/10.1111/gcb.16903.

[9] Kelland, Mike E., et al. "Increased Yield and CO 2 Sequestration Potential With the C 4 Cereal Sorghum Bicolor Cultivated in Basaltic Rock Dust‐Amended Agricultural Soil." Global Change Biology, vol. 26, no. 6, Apr. 2020, pp. 3658–76, https://doi.org/10.1111/gcb.15089.

[10] de Oliveira Garcia, Wagner, et al. "Impacts of Enhanced Weathering on Biomass Production for Negative Emission Technologies and Soil Hydrology." Biogeosciences, vol. 17, no. 7, Apr. 2020, pp. 2107–33, https://doi.org/10.5194/bg-17-2107-2020.

[11] Luczak, Joshua. "Concerns and Questions About Carbon Dioxide Removal Technologies." WIREs Climate Change, vol. 17, no. 3, May 2026, https://doi.org/10.1002/wcc.70063.

[12] Lawford-Smith, H., and A. Currie. "Accelerating the Carbon Cycle: The Ethics of Enhanced Weathering." Biology Letters, vol. 13, no. 4, Apr. 2017, p. 20160859, https://doi.org/10.1098/rsbl.2016.0859.

[13] Hartmann, Jens, et al. "Enhanced Chemical Weathering as a Geoengineering Strategy to Reduce Atmospheric Carbon Dioxide, Supply Nutrients, and Mitigate Ocean Acidification." Reviews of Geophysics, vol. 51, no. 2, Apr. 2013, pp. 113–49, https://doi.org/10.1002/rog.20004.

[14] Köhler, Peter, et al. "Geoengineering Impact of Open Ocean Dissolution of Olivine on Atmospheric CO2, Surface Ocean pH and Marine Biology." Environmental Research Letters, vol. 8, no. 1, Jan. 2013, p. 014009, https://doi.org/10.1088/1748-9326/8/1/014009.

[15] Montserrat, Francesc, et al. "Olivine Dissolution in Seawater: Implications for CO2Sequestration Through Enhanced Weathering in Coastal Environments." Environmental Science & Technology, vol. 51, no. 7, Mar. 2017, pp. 3960–72, https://doi.org/10.1021/acs.est.6b05942.

[16] McDonnell, Todd C., et al. "ForestERW: A Spatially Explicit Carbon Dioxide Removal Model for Practical Enhanced Rock Weathering

Applications in United States Forests." Frontiers in Climate, vol. 8, Aug. 2026, https://doi.org/10.3389/fclim.2026.1855001.

[17] Danczyk, Megan, and Christopher Oze. "Suitability of Rocks, Minerals, and Cement Waste for CO2 Removal via Enhanced Rock Weathering." Communications Chemistry, vol. 7, no. 1, Nov. 2024, https://doi.org/10.1038/s42004-024-01361-6.

[18] Maxbauer, Daniel P., et al. "Evidence for Carbon Dioxide Removal via Enhanced Rock Weathering With Steel Slag, Though Not Basalt, in a Midwestern U.S. Field Trial." Frontiers in Climate, vol. 7, Jan. 2026, https://doi.org/10.3389/fclim.2025.1657058.

[19] Baek, Seung H, et al. "Impact of Climate on the Global Capacity for Enhanced Rock Weathering on Croplands." Earth's Future, vol. 11, no. 8, Aug. 2023, https://doi.org/10.1029/2023ef003698.

[20] Beerling, David J., et al. "Farming With Crops and Rocks to Address Global Climate, Food and Soil Security." Nature Plants, vol. 4, no. 3, Feb. 2018, pp. 138–47, https://doi.org/10.1038/s41477-018-0108-y.

[21] Kantola, Ilsa B., et al. "Potential of Global Croplands and Bioenergy Crops for Climate Change Mitigation Through Deployment for Enhanced Weathering." Biology Letters, vol. 13, no. 4, Apr. 2017, p. 20160714, https://doi.org/10.1098/rsbl.2016.0714.

[22] Debona, Daniel, et al. "Silicon's Role in Abiotic and Biotic Plant Stresses." Annual Review of Phytopathology, vol. 55, no. 1, Aug. 2017, pp. 85–107, https://doi.org/10.1146/annurev-phyto-080516-035312.

[23] EPA. "Effects of Ocean and Coastal Acidification on Ecosystems." US EPA, 14 Nov. 2018, https://www.epa.gov/ocean-acidification/effects-ocean-and-coastal-acidification-ecosystems.

[24] Hangx, Suzanne J. T., and Christopher J. Spiers. "Coastal Spreading of Olivine to Control Atmospheric CO2 Concentrations: A Critical Analysis of Viability." International Journal of Greenhouse Gas Control, vol. 3, no. 6, Dec. 2009, pp. 757–67, https://doi.org/10.1016/j.ijggc.2009.07.001.

[25] Ogunkunle, Clement Oluseye, and Paul Ojo Fatoba. "Contamination and Spatial Distribution of Heavy Metals in Topsoil Surrounding a Mega Cement Factory." Atmospheric Pollution Research, vol. 5, no. 2, Apr. 2014, pp. 270–82, https://doi.org/10.5094/apr.2014.033.

[26] NOAA. "Ocean Acidification." National Oceanic and Atmospheric Administration, U.S. Department of Commerce, 25 Feb. 2025, https://www.noaa.gov/education/resource-collections/ocean-coasts/ocean-acidification.

[27] Lewis, Amy, et al. "Effects of Mineralogy, Chemistry and Physical Properties of Basalts on Carbon Capture Potential and Plant-Nutrient Element Release via Enhanced Weathering." Applied Geochemistry, vol. 132, June 2021, p. 105023, https://doi.org/10.1016/j.apgeochem.2021.105023.

[28] Rossato, Luciana, et al. "Impact of Soil Moisture on Crop Yields Over Brazilian Semiarid." Frontiers in Environmental Science, vol. 5, Nov. 2017, https://doi.org/10.3389/fenvs.2017.00073.

[29] Bissonnais, Yves Le, and Michael J. Singer. "Crusting, Runoff, and Erosion Response to Soil Water Content and Successive Rainfalls." Soil Science Society of America Journal, vol. 56, no. 6, Nov. 1992, pp. 1898–903, https://doi.org/10.2136/sssaj1992.03615995005600060042x.

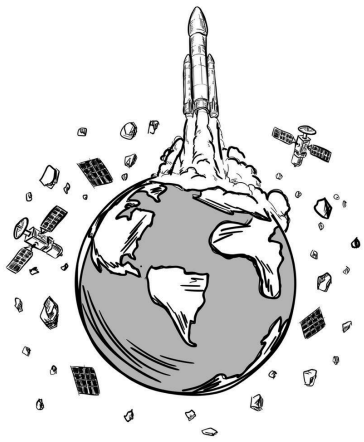
[30] Kozak, Catherine. "Vesta Says Olivine Sand Carbon Project at Duck Yielding Data | Coastal Review." Coastal Review, 13 Sept. 2024, https://coastalreview.org/2024/09/vesta-says-olivine-sand-carbon-project-at-duck-yielding-data/.

[31] "Duck, NC." Vesta.Earth, 2024, https://www.vesta.earth/duck.
[32] Pan, Shu-Yuan, et al. "CO2 Mineralization and Utilization Using Steel Slag for Establishing a Waste-to-Resource Supply Chain." Scientific Reports, vol. 7, no. 1, Dec. 2017, https://doi.org/10.1038/s41598-017-17648-9.
[33] Manning, David A. C. "Enhanced Rock Weathering — a Nature-Based Solution for Climate Mitigation." Green Energy and Sustainability, Mar. 2025, https://doi.org/10.47248/ges2505020003.
[34] Mowbray, Sean. Calls for Caution as Enhanced Rock Weathering Shows Carbon Capture Promise. Oct. 2024, https://doi.org/10.66709/news-289361.
[35] Oppon, Eunice, et al. "Sustainability Performance of Enhanced Weathering Across Countries: A Triple Bottom Line Approach." Energy Economics, vol. 136, June 2024, pp. 107722–22, https://doi.org/10.1016/j.eneco.2024.107722.
[36] Arnalds, Olafur, et al. "The Icelandic Volcanic Aeolian Environment: Processes and Impacts — a Review." Aeolian Research, vol. 20, Mar. 2016, pp. 176–95, https://doi.org/10.1016/j.aeolia.2016.01.004.
[37] Iqbal, Qaiser, et al. "Marble Dust Effect on the Air Quality: An Environmental Assessment Approach." Sustainability, vol. 14, no. 7, Mar. 2022, p. 3831, https://doi.org/10.3390/su14073831.
[38] Sager, Manfred. "Urban Soils and Road Dust—Civilization Effects and Metal Pollution—a Review." Environments, vol. 7, no. 11, Nov. 2020, p. 98, https://doi.org/10.3390/environments7110098.
[39] Schenker, Marc B. "Inorganic Agricultural Dust Exposure Causes Pneumoconiosis Among Farmworkers." Proceedings of the American Thoracic Society, vol. 7, no. 2, May 2010, pp. 107–10, https://doi.org/10.1513/pats.200906-036rm.
[40] Farmer, Andrew M. "The Effects of Dust on Vegetation—a Review." Environmental Pollution, vol. 79, no. 1, 1993, pp. 63–75, https://doi.org/10.1016/0269-7491(93)90179-r.
[41] Prospero, Joseph M., et al. "Characterizing the Annual Cycle of African Dust Transport to the Caribbean Basin and South America and Its Impact on the Environment and Air Quality." Global Biogeochemical Cycles, vol. 28, no. 7, July 2014, pp. 757–73, https://doi.org/10.1002/2013gb004802.
[42] Poinen-Rughooputh, Satiavani, et al. "Occupational Exposure to Silica Dust and Risk of Lung Cancer: An Updated Meta-Analysis of Epidemiological Studies." BMC Public Health, vol. 16, no. 1, Nov. 2016, https://doi.org/10.1186/s12889-016-3791-5.
[43] Wu, Shixian, et al. "Enhanced Dust Removal via the Synergy of a Standing Wave Acoustic Field and High-Pressure Spray: An Integrated Experimental and Numerical Study." PLOS One, edited by Antonio Riveiro Rodríguez, vol. 21, no. 2, Feb. 2026, p. e0339045, https://doi.org/10.1371/journal.pone.0339045.
[44] Dupla, Xavier, et al. "Potential Accumulation of Toxic Trace Elements in Soils During Enhanced Rock Weathering." European Journal of Soil Science, vol. 74, no. 1, Jan. 2023, https://doi.org/10.1111/ejss.13343.
[45] Fagnano, Massimo, et al. "Copper Accumulation in Agricultural Soils: Risks for the Food Chain and Soil Microbial Populations." Science of The Total Environment, vol. 734, Sept. 2020, p. 139434, https://doi.org/10.1016/j.scitotenv.2020.139434.
[46] Adrees, Muhammad, et al. "Mechanisms of Silicon-Mediated Alleviation of Heavy Metal Toxicity in Plants: A Review." Ecotoxicology and Environmental Safety, vol. 119, Sept. 2015, pp. 186–97, https://doi.org/10.1016/j.ecoenv.2015.05.011.
[47] Suhrhoff, Tim Jesper. "Phytoprevention of Heavy Metal Contamination From Terrestrial Enhanced Weathering: Can Plants Save the

Day?" Frontiers in Climate, vol. 3, Jan. 2022, https://doi.org/10.3389/fclim.2021.820204.
[48] Foteinis, Spyros, et al. "Life Cycle Assessment of Coastal Enhanced Weathering for Carbon Dioxide Removal From Air." Environmental Science & Technology, vol. 57, no. 15, Apr. 2023, pp. 6169–78, https://doi.org/10.1021/acs.est.2c08633.
[49] Hutchins, David A., et al. Responses of Globally Important Phytoplankton Species to Olivine Dissolution Products and Implications for Carbon Dioxide Removal via Ocean Alkalinity Enhancement. Apr. 2023, https://doi.org/10.1101/2023.04.08.536121.
[50] Dupla, Xavier, Romane Claustre, et al. "Let the Dust Settle: Impact of Enhanced Rock Weathering on Soil Biological, Physical, and Geochemical Fertility." Science of The Total Environment, vol. 954, Sept. 2024, p. 176297, https://doi.org/10.1016/j.scitotenv.2024.176297.
[51] O'Sullivan, Kate, et al. "Who Pays for Carbon Dioxide Removal? Public Perceptions of Risk and Fairness of Enhanced Rock Weathering in the UK." Humanities and Social Sciences Communications, vol. 12, no. 1, July 2025, https://doi.org/10.1057/s41599-025-05384-9.
[52] Wong, Pak-Hang. "Maintenance Required: The Ethics of Geoengineering and Post-Implementation Scenarios." Ethics, Policy & Environment, vol. 17, no. 2, May 2014, pp. 186–91, https://doi.org/10.1080/21550085.2014.926090.
[53] Heyward, Clare. "Benefiting From Climate Geoengineering and Corresponding Remedial Duties: The Case of Unforeseeable Harms." Journal of Applied Philosophy, vol. 31, no. 4, July 2014, pp. 405–19, https://doi.org/10.1111/japp.12075.
[54] Levy, Charlotte R, et al. "Enhanced Rock Weathering for Carbon Removal–Monitoring and Mitigating Potential Environmental Impacts on Agricultural Land." Environmental Science & Technology, vol. 58, no. 39, Sept. 2024, pp. 17215–26, https://doi.org/10.1021/acs.est.4c02368.

Journal of Ethics in Scientific and Technological Innovations

The Space Race's Second Lap! An Ethical Discussion Regarding the Modern Space Revitalization



Gavin Finkel[1]
[1]*HAFTR High School, Cedarhurst, NY 11516*

Abstract

With the rise and recent success of private space-oriented companies over the past few years like SpaceX and Blue Origin along with the US' rejuvenated efforts to explore the moon through the Artemis program, space travel has been at the forefront of scientific discourse only comparable to the first Space Race during the Cold War in the 1960s. However, in comparison, the modern day "race" occurring in the US has a significantly higher focus on the potential economics and settlement of outer space. The shifted players and focus of the current space-related excitement calls for the assessment of the ethical implications surrounding the topic. Now, not only government organizations but also private companies are participating in the "race" to achieve space innovation. This is coupled with new motivations to traverse space such as militarily, for the construction of a moon and Mars bases, and to support the satellite and tourism industries. All the while, space junk has been accumulating in the atmosphere and rocket launches have launched hundreds of tons of carbon dioxide into the atmosphere. However, morality in the US' investment of space heavily depends on one's own vantage point and how much one values rewards that will only arrive decades to come. At the end of the day, space has always been an investment; only now, the economic gain, military reinforcement, and potential for settlement on celestial bodies appears closer than ever before. Nevertheless, the US should push initiatives that hold private space companies accountable, limit space junk, and decrease negative climate impacts. Overall, this article seeks to answer the overarching question of "How should humanity approach space?" from a moral perspective, as both humanity's technological prowess and potential hubris brings us closer to the stars with each new innovation.

1. Introduction

Humanity has been researching the stars for centuries but only had the technological prowess to finally leave the Earth in the 1960s. This time period of immense space innovation, dubbed the Space Race, is crucial to understanding the context of modern day space ethics.

The first sparks of the space race occurred during the tail end of WWII when Nazi Germany completed their V-2 rocket program, revolutionizing what a rocket could be and accomplish in warfare due to the V-2 rocket's liquid propellant system and never-before-seen maximum travel range (200 miles) [1]. Thus, as

tensions between the United States and the Soviet Union increased, the two superpowers sought to develop increasingly efficient and stronger ballistic missile technology as a part of the larger competition to create the strongest and most innovative weapons [2]. This resulted in rocket technology becoming increasingly capable of carrying heavier loads across greater distances, thereby opening up the possibility to send these rockets farther than ever before- to space [3].

The US' and Soviet Union's competition to innovate space-related technologies, or the Space Race, was largely motivated purely for the sake of political superiority over the opposing country [2]. The conflict only really began in 1957 when the Soviet Union launched the first-ever satellite, Sputnik, before the US, which announced its goal for space-related endeavors in 1955. The Space Race became a symbol of the two countries' technological capabilities and national pride during the Cold War as both of their governments each attempted to cross space-related milestones before the other, ultimately culminating in the announcement of the US' goal to send astronauts to the moon's surface in 1961.

Over the course of the race (1955-1969), the Soviet Union succeeded in the majority of space endeavors such as launching the first satellite, man, and woman to space [2]. Regardless of these feats, in 1969 through the Apollo 11 rocket, the US pulled through on the greatest hurdle—sending man to the moon, cementing the US as the major space world power [2]. The US was able to accomplish this historic task by contracting thousands of international scientists to aid its newly formed extraterrestrial agency [4], the National Aeronautics and Space Administration (NASA), and by using NASA to develop incremental improvements and innovations to its single-use (were discarded after liftoff) "Saturn V" rockets, which contributed to the final formation of Apollo 11 [5].

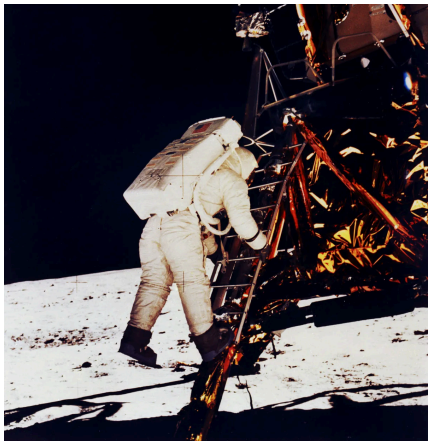


Fig 1. Neil Armstrong steps out of Apollo 11 and becomes the first human to walk on the moon.[6]

However, as a result of the US' victory, the nationwide space obsession consequently came to an end. Satisfied with Apollo 11, the American people and many politicians felt the US had won the Space Race and therefore space was not valuable enough to continue investment [7]. Space exploration never generated capital and after Apollo 11 not even much patriotism anymore. Thus, also faced with the heavily expensive Vietnam War, President Richard Nixon decreased NASA's budget in order to funnel that capital to matters deemed of a higher priority. Additionally, in 1970 the Apollo 13 rocket failed midflight, causing a near fatal incident to the mission's astronauts. The lower budgeting and all around decreased morale contributed to the dramatic downscaling of NASA's operations post-1969, demonstrated by the organization's cancellation of three of its future Apollo programs.

Furthermore, the deprioritization of space was compounded when in 1975 the US and Soviet Union collaborated on the Apollo-Soyuz Mission, which involved American and Russian astronauts shaking hands in space [7]. Then, between 1984 and 2011,

America, Canada, Europe, Japan, and even Russia - the successor to the Soviet Union - all contributed to the construction of the International Space Station (ISS) [8]. These two events provided a satisfying end to the Space Race, concluding the decades-long space feud with international peace and cooperation. As a result, space exploration was not seen as critical of an endeavor anymore. This is evident in NASA's 2011 goals that highlighted many concepts aside from space exploration such as understanding the deep cosmos, cooperation, and a variety of Earth-related activities [9].

With the past concluded, how did the present become the Space Race's "Second Lap"? Modern space revitalization in the past decade has been brought upon by lowered launch costs from improved rockets, the success of private space companies, computing innovations [10], and new political rivalries [11]. These factors have contributed to the space industry reaching $613 billion spent in 2024, a monumental record [12].

However, with all of that context in mind, space has been undergoing very major changes since the 1960s. The industry has shifted from regulated government operations to accepting private companies who prioritize innovation as much as turning a profit [13], the motivations have shifted from Cold War patriotism and political superiority [2] to economic [14] and ambitious interplanetary causes [15], and the single-use rockets like in the Apollo program along with the remains of dead satellites are still abandoning hundreds of billions of dollars not just of discarded parts but of human time, effort, and tax dollars that are left to float in space indefinitely [16]. Therefore, this article seeks to answer the question of "How to approach space?" from a moral perspective, contrasting the previous Space Race with the Second Lap to examine the morality of the new who, why, where, and how of space exploration.

2. Who: Private Companies and Government

One of the most notable differences between the 1960s and today is the decreased reliance of government operations to fuel space innovation.

The Cold War vitriol and competitive atmosphere pushed N.A.S.A. into the forefront with a complementary gargantuan budget—totalling at maximum 4.5% of government annual spending [17].

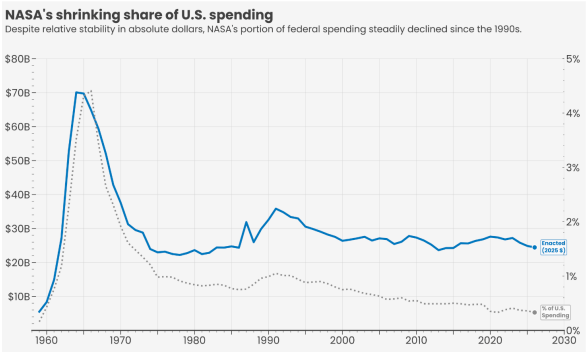


Fig 2. US budget allocated to N.A.S.A since inception, peaking during the heights of the Space Race[17]

With the budget to do so, NASA prioritized results rather than efficiency in order to compete with the Soviet Union [18]. Therefore, in theory, the success of private companies should be a positive aspect of space innovation in the past decade. Private, non-government owned space companies are forced to work on stricter budgets that must prioritize cost-cutting measures and technological innovation rather than pure results [19]. Furthermore, these private space companies have not only been innovating but also have been successful in their endeavors. With the rise of the satellite industry, the private company SpaceX has been a dominant force in this field through their reusable Falcon-9 rocket, being attributed to launching 75% of payload into orbit [20]. The private company Blue Origin has similarly seen success by tapping into the space tourism industry and, as of 2026, has sent

98 people into space over the course of 38 launches [21].

However, in all of human history, this is really the first time private space companies have existed and been so successful; as a result, the moral concerns must be highlighted as well. For instance, NASA has spent $15 billion worth of contracts sourced from taxpayer money to SpaceX [13], but how much of that money is truly going toward innovation? In a private company, that funding is sent not only to the workers but the stakeholders invested in the company as well. Do the American people truly want their taxpayer money to go to millionaires and billionaires? Compare that to NASA in the 1960s where the Apollo programs supplied taxpayer money directly to around 400,000 NASA workers [13] but at the expense of efficiency and innovation [18]. Therefore, there appears to be a moral tradeoff: greater innovation from market competition but with taxpayer money dispersed less to those who deserve it versus less pressure to innovate but with taxpayer money clearly being funneled to the workers in the company.

Furthermore, caution to the government funding private companies is especially relevant when examining NASA's previous collaboration with Boeing. As of February 2025, Boeing has lost a staggering $2 billion in its contract with NASA to construct "Starlink" or a system to easily transport crews back and forth to the ISS [22]. This failure was indeed committed under NASA's "Commercial Crew Program"—an initiative focused on transporting crews to the ISS via funding of private industry [23]. SpaceX was involved in the same program but with significantly superior results, obtained through SpaceX's Falcon 9 reusable rocket. So, in a way, government funding to private companies contributes to society by enlightening the public and the government on which companies can be relied upon for future government aid. Additionally, there should be little fears of NASA's targeted funding creating monopolies as the administration, through its Private Sector Engagement Office, goes out of its way to make connections with new private companies by allocating funding to their programs and sharing NASA data [24].

Therefore, if private companies are going to be receiving contracts then transparency must be preserved. When Americans saw the failure of Apollo 13 and the near-death of its crew, a reasonable backlash occurred that became a cause for the downturn of public perception towards NASA following the US' victory in the Space Race [7]. On the other hand, SpaceX had two launches that had forced reschedulings due to problematic rocket parts within one week of each other (July 16 and 20 in 2026) yet had no obligation to inform the public of these events [25]. Especially since SpaceX has received government funding for a variety of projects, such as transportation flights to the ISS, aid in Artemis's return to the moon, and sending military satellites into orbit [26], it is imperative that the results of these contracts are not forgotten and instead shared with the American public.

The great effect of private companies on space exploration is undeniable—but that is not to say the government has lost its role. The existence of NASA allows the space industry to operate within a moral framework that is trusted to serve the interests of humanity. If a private company sought to develop a moon base independently, nothing would be able to directly limit their capabilities if immoral or anti-humanity actions and business decisions would be pursued. On the other hand, NASA funding a private company for the same reason assumes that the base will be constructed for the betterment of American society [27] and, if that is not the case, can pull the base's funding. Furthermore, on the topic of a moon base, government operations are also critical to subsidizing gigantic projects too risky to be

tackled by private companies. Famously, the ISS cost approximately $150 billion to fully develop, making the station the single most expensive object humanity has built [28]. The ISS has contributed to over 3,000 space experiments from astronauts across 108 countries and would not exist if not for vast government funding [28].

Regardless of NASA pushing the boundaries of what is possible in space, the organization is not out of the clear. This idea was most infamously demonstrated by NASA's Space Shuttle. The Space Shuttle was a program which sought to construct a reusable rocket that could launch back and forth from space with little cost and short turn-around time [29], reminiscent in hindsight of SpaceX's reusable Falcon 9 rocket. Nevertheless, the shuttle ended up being more expensive ($450 million per launch) and had a higher turn-around (88 days) than originally planned. Sadly, out of the six shuttles, two of them (the Challenger in 1986 and Columbia in 2003) combusted and resulted in the deaths of 14 people; these tragedies were attributed to lack of safety precautions and fundamental failures within NASA's organizational culture [30].



SPECIAL REPORT ON SHUTTLE DISASTER/9 PAGES

San Francisco Chronicle

The Largest Daily Circulation in Northern California

122nd Year No. 11 ★★★★★ WEDNESDAY, JANUARY 29, 1986 777-1111 25 CENTS

WHAT WENT WRONG

7 Killed In Shuttle Explosion

By Boyce Rensberger
Washington Post

Cape Canaveral

The space shuttle Challenger, carrying a crew that included schoolteacher Christa McAuliffe, exploded in a burst of flames 74 seconds after lift-off yesterday, killing all seven aboard and stunning a world made witness to the event by television.

The unexplained explosion occurred at 11:38 a.m. EST as the flight seemed to be proceeding flawlessly at about 2000 feet per second, 10 miles above Earth and eight miles downrange from Cape Canaveral.

The spacecraft disintegrated into bits of debris that rained into the Atlantic Ocean. Those aboard, still strapped into their seats, had no escape system.

It was the worst accident in the history of space exploration and the first time anyone has been killed during an American space flight.

The tragedy occurred 10 years and one day after U.S. astronauts Virgil (Gus) Grissom, Roger Chaffee

Back Page Col. 5

2 Areas on Fuel Tanks Suspected

By Thomas H. Maugh II
Los Angeles Times

Cape Canaveral

A little more than a minute after the space shuttle Challenger lifted off from pad 39B at the Kennedy Space Center yesterday morning a thin tongue of flame appeared between the booster rockets and the shuttle's main fuel tank.

The flame was not seen by observers on the ground, but slow motion videotapes of the launch show it clearly.

Less than a second later, nearly 200,000 pounds of liquid hydrogen in the main fuel tank exploded in a massive fireball that enveloped the shuttle and the two solid fuel rocket boosters flying off in near opposite directions. The explosion destroyed the orbiter.

Although there is little doubt about what happened, exactly how it happened remained a major mystery. Yesterday, however, most of the speculation centered on two possibilities.

One is that a defect in the solid rocket booster caused exhaust

Back Page Col. 1

Smoke plumes trailed the solid fuel rockets (upward, right) as they continued skyward after the liquid fuel tank and shuttle exploded

Fig 3. Article by the San Francisco Chronicle detailing the tragic failure of N.A.S.A.'s Challenger Shuttle[31]

The Space Shuttle program was then cancelled due to its high costs, slow turnaround, and blatant safety problems [29]. Therefore, under the continued assumption that the most moral government space exploration programs are achieved through maximizing the efficient use of dispersed taxpayer money in order to optimize innovation and improve results, the combination of both government and private companies maintaining a relationship and shared utilization of taxpayer money appears to be the most moral option. This system maximizes efficiency by letting government operations push the ceiling of space exploration all the while private companies make innovations cheaper and, in the case of the Falcon 9 (which had a cost of $74 million per launch [32] with a turn-around approximately between 9 days and two months [33]) versus the Space Shuttle, even superior.

3. Why and to Where: Examining Motivations

As the players for the Space Race have changed, the motivations for space exploration have shifted as well. During the Cold War, only the US and Soviet Union had the economic ability to invest in space, naturally leading to space largely being focused on innovation for the sake of perceived political greatness or espionage [2]. Today, NASA works to innovate its scientific and technological capabilities [34], improve its manned missions to the Moon and later Mars [14], and curb the possibility of the US falling behind China's fledgling but successful space programs, which are occasionally in partnership with Russia. On the other hand, private companies, by definition, operate their space programs with the clearer goal of economic incentives which has manifested itself through the satellite industry [14] and space tourism [35]. All of these causes require moral examination of whether their economic or political sources outweigh the high

costs and risks associated with their implementation.

3.1. Millitary

To preface, political or military-based space innovation is in higher demand now than ever before. Even during the Cold War when space technology was in its infancy, the US army benefitted from the US' top space program, developing satellites that were able to take pictures of the Soviet Union from outer space for reconnaissance, specifically to inspect the strength of Soviet firearms [36]. Thus, certainly today when the world has significantly greater technology and when China and Russia are partnering to expand their space programs, space for military operations may become essential. Most notably, as of 2026, the current US administration is developing the "Golden Dome" project which seeks to act as a defensive barrier for the US against enemy missiles using satellites to track when those missiles are fired [37]. Due to the increasing relevance of missiles in warfare [38], it is natural for countries to need to develop anti-missile technology. Thus, the moral issue lies more in offensive space projects, such as firing missiles or projectiles fired from satellites in orbit. If missiles could be launched from orbit, the natural next step to optimize the destructive potential of those satellites is to prepare a nuclear bomb in space. Nevertheless, that likely would not change much because a nuclear bomb that takes, for example, 30 minutes versus five when assisted by space technology to spark nuclear fallout does not make much of a difference. Thus, speculatively, the only difference an orbital strike would cause is increasing the power gap between strong and weaker countries. An army capable of leveling a country would be able to accomplish that task in merely less time with space technology while a country unable to defend against missiles would simply become that much more helpless. From a different perspective, offensive space technologies also have the chance to affect a completely different sector of geopolitical tensions. This is due to the highly experimental "God's Rods" program which seeks to fire high velocity tungsten rods from satellites in outer space that could act as a more resource-efficient alternative to missiles [39].

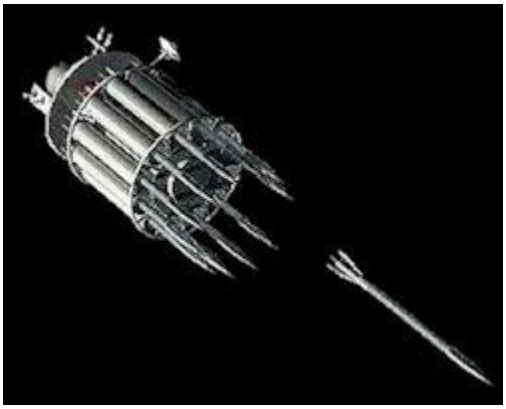


Fig 4. Mock-up of what the God's Rods program could resemble if developed[40]

Additionally, the US' Department of Defense has considered the possibility of establishing satellites with laser capabilities in space, demonstrating another missile alternative [41]. Missiles are developed using rare minerals that are predominantly supplied by the US' rival, China [42], while the tungsten rod projectiles utilize tungsten which can be sourced from a variety of states in the US [43]. Therefore, the development of tungsten-based space weaponry or a space laser system could aid in establishing the US' continuous status as a world power in the future by decreasing military reliance with China. 72% of Americans in a poll of 2,500 in 2024 who all agreed that the US "should play a lead or major role in world affairs" [44]. Hence, if a government's role is to serve and protect its people, then further exploration into the field of space weaponry can be seen as the moral option because it can cement the US' future military and thus political superiority. In sum, space defensive programs are an ethical necessity

while the morality of offensive space strikes at the moment appears to be more so reliant on one's preconceived notions due to the concept merely speeding up or increasing the efficiency of already militarily strong countries with the potential benefit of decreasing trade reliance.

3.2. The Moon and Mars

Returning to space exploration, the most popular discussion of the topic, currently, regards the renaissance of Moon exploration, championed by NASA which is collaborating with Blue Origin and SpaceX through the "Artemis" mission [15]. An 2007 article by the Nationals Academies for Sciences, Engineering, and Math requested by NASA discusses the scientific developments that can arise from a moon base [45]. For example, improved telescope technology and a greater understanding of gravity; ancient life and evolution; Earth's climate, ionosphere, and plasmosphere; the Sun; and extrasolar planets. However, aside from only the research on Earth's climate, all of these developments are quite abstract and do not appear to directly benefit humanity. Many of these new opportunities for research would be incredible discoveries but do they warrant the current estimated costs of $30 billion for a moon base [46]? NASA's own website concerning the moon base is vague, lacking exact reasons relevant to humanity's benefit from a moon base. A short blurb claims that a moon base would "enable increasingly ambitious scientific research, technology demonstrations, and exploration," but the page is less focused on the moon base itself but more on the innovation of future space exploration that will arise from it: missions to Mars [47]. The concept of the main driving force of moon exploration being a stepping stone to Mars is similarly expressed by SpaceX's website [48]. Therefore, in order to establish the ethics of spending $30 billion on a moon base, exploring whether humanity would benefit from a Mars base is crucial to fully understanding the situation.

Though Mars will not automatically catalyze wealth and innovation that can aid life back on Earth, the benefits of Mars travel are substantially clearer than those of Moon travel. First of all, the ability for long-term inhabitance on Mars is significantly greater than on the Moon. This is due the fact that, unlike the moon, Mars contains elements necessary for life and human existence—such as carbon, nitrogen, hydrogen, and oxygen—along with a water supply that is theorized to be equal to an ocean 100 meters deep that covers the whole planet [49]. Furthermore, Mars has many opportunities for self-sufficient energy systems like wind and solar power; wind power can be sourced from Mars' active weather and solar power from Mars' large carbon and hydrogen supply that can be used to make solar panels [49]. More importantly, Mars can give back to humanity through non-theoretical-based and economic pathways. The planet possesses valuable minerals that are useful in industry and that are rare on Earth, along with having a closer proximity to asteroids that are similarly notable for containing rare minerals [49]. Outside of minerals, Mars is the only other planet in the Solar System outside of the Earth that has the potential for agriculture [49]. Lastly, Mars has a large supply of deuterium, a material useful for nuclear fission, which is five times the size of Earth's [49]. All of these materials—rare minerals from the planet and asteroids, crops, and deuterium—can be sent back to Earth and therefore give a Mars base proper economic motivation.

However, the ethical queries of funding billions and billions of dollars towards Mars do not dissipate if the program is economically valuable because the conditions on Mars are extreme and may be dangerous to potential astronauts. The planet has an average yearly temperature of -60° Celsius (-76° Fahrenheit)

with an ultraviolet radiation flux that damages DNA over a thousand times more than to the surface of Earth [50]. Additionally, Mars lacks a source of water on its surface, has an atmosphere that is obviously unbreathable, and soil that is potentially toxic [50]. Furthermore, the NASA website details the variety of issues related to a years-long Mars exploration, such as the radiation that has a greater effect on human health than on Earth; lower gravity that causes a decrease of approximately 1-1.5% mineral density in astronauts' bones per month, which may be permanent; loss of muscle mass that is faster than on Earth, requiring strict diets and exercise routines to balance out; and increased risk of kidney stones and dehydration [51]. Even further, living on Mars would involve prolonged isolation and confinement in a closed environment, causing elevated stress hormones and alterations to the immune system which could then increase susceptibility to allergies or certain illnesses [51].



Fig 5. Mock-up of a theoretical Mars base; notice how narrow and tight the structure is.[52]

To summarize, Mars has a greater potential for long-term inhabitants than the Moon and clear financial incentives to aid humanity on Earth through the nuclear fission and materials industries. On the other hand, Mars exploration is associated with significant health risks that require extreme precautions to the astronauts who would be living in a physically and mentally unenviable environment. However, a newfound space trade industry can strengthen the US' access to important minerals and thereby allow the US to thrive even after some of Earth's mineral resources run out, which is estimated to occur as soon as 2050 [53]. Therefore, combined with the fact that NASA is attempting to alleviate concerns regarding the wellbeing of astronauts [51], exploration to Mars may genuinely benefit humanity due to its aforementioned meaningful opportunities. As a result, a moon base indeed has merit by providing a stepping stone to future space exploration and industry potentially even further than Mars.

3.3. Profit

The final significant incentive to space is the rapidly growing nearly purely economic motivations—the development of satellites and tourism. Satellites are especially relevant in this discussion because they have provided an economic incentive for space strong enough to fuel the flames of success for private space companies, launching the Second Lap [14]. The significance of satellites is evident in that GPS satellites alone have generated $1.4 billion in economic benefits as of 2020 [14] and that 2025 had the record number of objects sent to space, with 4,510 objects being sent out of a relatively small 25,000 objects in total that have been sent to space [54].

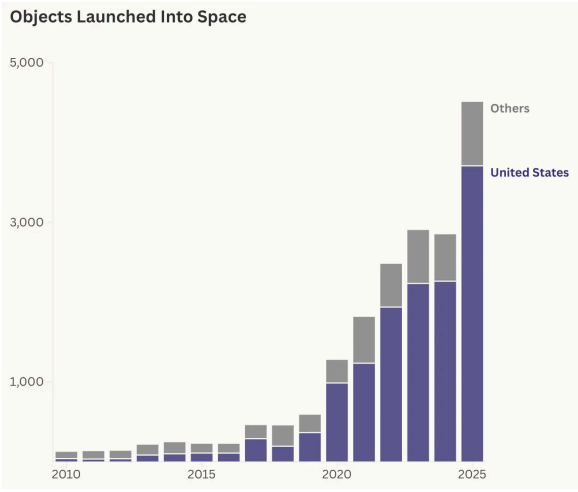


Fig 6. Number of objects launched to space from 2010-2025[54]

Furthermore, satellites have also been incredibly beneficial to humanity by providing faster communication, navigation and positioning technologies; climate examination and prediction; and large-scale data transfers [55]. Additionally, the number and capabilities of satellites are increasing with SpaceX's Starlink program, which is able to provide anyone on Earth's surface with reliable Wi-Fi no matter how remote the location (for 55$ a month) [56]. On the tourism side, both SpaceX and Blue Origin have invested in successful space tourism programs [57]. Blue Origin has been more so invested in this endeavor, sending all types of people from different backgrounds, races, ethnicities, gender, and body types to space through their New Shepard program [21]. Diversity in space is not a simple feat and has required decades of innovation; the first Apollo programs required men below 5'11'' who were military test pilots (and therefore were all white) [58]. Even though satellites do provide a better living experience on Earth, the motivation for satellites are largely economically motivated and have veered from space's original goal to provide societal benefit. Nevertheless, private space companies, as discussed, are at the forefront of space innovation just like NASA, and space innovation from any company or sector can be shared with others and is beneficial to the industry in the long term [24]. Therefore, these companies can be allowed to try to make a profit.

4. How: The "Travel" in "Space Travel"

Now that the players, background, and goals have been examined, all that is left to discuss is how space exploration actually occurs: the rocket.

4.1. Single-use Rockets

Beginning in the Apollo-era, rockets were designed to be one-use [5] but mostly because of technological limitations and NASA's gargantuan budget [18]. If you had such a large capital combined with the fact that rockets were still a practically brand-new technology, why spend money on multi-use rockets, an invention that implies the necessity of cost saving, when the Soviets were succeeding with simple single-use ones? As such, this fact led to the Apollo rockets which were designed to discard their Saturn V segment, a gigantic feat of engineering that was 363 feet tall and weighed 6.2 million pounds [5]. A total of 13 Saturn V's were launched [5] and consequently discarded, abandoning a total of 80,600,000 pounds of material. This material left to float and accumulate in space is aptly referred to as "space junk". NASA, as discussed, only began a multi-use rockets program a couple of decades later through the Space Shuttle program [29], yet the damage was done and a continuous precedent of constructing single-use rockets regardless of the space junk they discard had been set.

How has this worked out practically? As of 2023, a calculated between $570 billion and $1.2 trillion of scrap metal, or a staggering 5,312 to 19,124 tonnes, has been lost in space [16]. To emphasize, in 2022 the average automobile had a weight 4,094 pounds [59], so there is the equivalent weight of 11,711,000 cars to the amount of space junk floating in space, according to the minimum estimated by the previous calculation. This practice is not sustainable for long-term space use as it wastes the metals, alloys, and other materials that are necessary for space infrastructure [60]. Furthermore, from an environmental perspective, rockets already emit around 200 to 300 tons of carbon dioxide per launch [61]. Therefore, due to the space industries' intertwined high emission count, having to then

make a brand new rocket sends a staggeringly higher amount of carbon dioxide into the atmosphere per launch relative to a reusable rocket that does not require an entirely or nearly entirely new rocket per launch. Though it is difficult to calculate the exact emissions released from the construction of a rocket, rocket materials like titanium and aluminum, which are both heavily used for the vehicle's structural properties, are high emitters [62]. Titanium releases 1.9 tons of CO2 per ton of its formation [63] while aluminum releases 1.5-2 tonnes per ton of its formation [64] (converting from the previous statistic, one Saturn V weighs 3,150 tons or approximately 2,858 tonnes). As the space industry becomes more and more significant, single-use rockets have great potential for environmental harm.

4.2. Alternative Rockets

But what if rockets did not release any carbon dioxide in the first place? Research has already been in development for such a rocket, with the most likely model being powered by nuclear energy [65] and lacks the carbon emissions prevalent in modern rockets [66]. These nuclear-powered rockets are also lauded for being twice as efficient and having greater speed compared to the current chemical reactors in rockets, theoretically lowering the time to reach Mars by a quarter of the estimated time using chemical rockets [66]. The aforementioned astronauts who have to struggle to stay alive for long periods of time in space would likely appreciate that statistic. Nevertheless, NASA for the time being states it will continue pursuing chemical fuel due to its technical and economical benefits but is still heavily considering a nuclear-powered flight for future trips to Mars [65]. On that note, NASA has also been researching an even more experimental version of space travel that requires no propellant at all [67]. Examples of these technologies for short space missions include a solar sail—a large sheet of material that uses solar radiation to generate thrust—or a space tether—a kilometer-plus long wire that floats in orbit and can be used to efficiently raise or lower a spacecraft into and out of orbit. Although some of these projects may seem far off, initiatives like these are important for conserving the limited resources on Earth and preserving the planet's climate.

For these similar reasons, reusable rockets like SpaceX's Falcon-9 appear as the next clear step forward. Certainly, reusable rockets do still release tons of emissions but comparing the emissions from a launch versus a launch with a whole new rocket is an improved situation. Also, simply limiting space programs to limit emissions is not a feasible solution due to, as previously discussed, space's status as an integral part of future military, economic, and satellite technology. So, should NASA end its single-use rocket program in order to preserve materials and lower emissions? Preferably yes, but reality is not that simple. The issue is that the current technology for rocket reusability is not scalable. The Falcon-9 rocket can hold enough fuel for a mission into orbit but not to, for instance, the moon because fuel requirements first must increase to reach the moon then have to even further increase in order to balance out the weight added from the original increase of fuel [68]. Hence, NASA's Artemis rockets require multiple heavy "SLS" boosters in order to accommodate their journey to the moon that then need to be eventually thrown away because they add too much weight to the rocket [69]. As a result, current reusable rockets are unable to travel farther than Earth's orbit lest they require high amounts of fuel that exponentially increase, which also may cause the fuel tanks to need to be discarded [70] and thereby not make the rocket reusable [68]. Regardless, reusable rockets are certainly the moral future of space in order to limit emissions and increase resource

efficiency, especially if the technology becomes more viable.

4.3. Space Junk

Though reusable rockets can curb the amount of human-created material thrown away in space [65], the rise of space junk as previously touched upon by single-use rockets is a growing issue in the modern space industry. Currently, there are around 34,000 objects that are larger than 10 centimeters in orbit and, more specifically, 3,000 inoperative satellites [71].

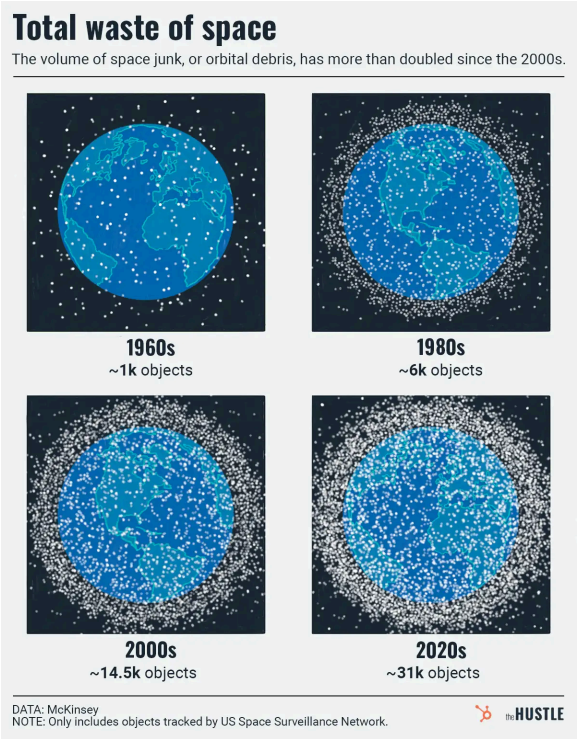


Fig 7. Graphical demonstration of the number of space junk orbiting the Earth since the 1960s[72]

So far, the ISS has been forced to adjust its course 25 times due to space junk since 1999 [71] and that similar risk will only increase as more and more satellites are launched into orbit [72]. To elaborate, there are fears that the amount of space junk will exponentially increase due to crashing satellites creating space junk, which in turn increases the chance of satellites hitting space junk, which in turn creates more space junk, and so on and so forth [72]. While space is in fact gigantic and at the moment space junk is not too large of a problem, the potential for future mid-orbit crashes is immense [71]. This heightened potential would theoretically force space projects to be rescheduled, consistently relocated in orbit, and be limited in scope and size as to risk a crash. Therefore, the United Nations has urged countries to remove dead satellites 25 years after their inoperation, yet this is not only difficult to enforce but also difficult to comply with as space-junk-removing-devices (for example, harpoons, nets, and lasers) are not commonplace and, even if they did exist, would struggle against larger or far away objects like discarded rocket thrusters. A possible and easy solution would be to simply wait for satellites to naturally drift and burn into Earth's orbit [72]. However, if space junk becomes such a prominent obstacle, waiting for the space junk to all return and disintegrate in Earth's atmosphere (which is also a waste of materials) may be too slow of a solution as the fall can take from weeks to many years to occur. Thus, for this problem only time will tell; the best solution at the current time appears to be to continue tracking space junk and invest in space-removing technologies that conserve material.

5. Discussion and Conclusion

The drastically changing US space industry has given rise to new moral queries and the need to reexamine old issues with a modern lens.

An ethical government ought to form the greatest productivity out of tax-payer money as possible in a benevolent and trustworthy fashion. The emerging private space sector does prioritize efficiency and innovation in space-related technologies, yet these companies should have an obligation of transparency to the American people if receiving federal funding

from NASA. On the other hand, NASA has had issues of helming inefficient projects, yet the association stands as a moral grounding for space innovation that also raises the ceiling of possibilities for space exploration missions. So, in order to optimize taxpayer and thus government morality, the current balance of the two sectors may be the most moral.

Moreover, due to space's danger but immense potential, motivations for space travel should be for the benefit of humanity as space travel risks human lives and is an extremely expensive initiative. However, following through on the political, interplanetary, and economic causes largely depends on one's preconceived notions of the significance of the US maintaining its role as a world power. For instance, defense spending on space is a moral necessity to protect a country from rising missile warfare while offensive spending merely increases the strength of already powerful countries. Similarly, spaceflight to other locations in the solar system, currently focusing on the Moon and Mars, can strengthen the innovation and pool of resources in the U.S., supporting the economy and certain industries but not directly the people, and that aid to the people is only expected to arrive in the far future once these space bases are firmly established. Furthermore, astronauts will have to go through the grueling process of participating in these missions but it may be worth their physical and mental sacrifices if it contributes to future prosperity. The economic motivations for space travel, though shallower, have pushed the industry toward new heights and allowed for its high productivity today.

Lastly, even if the causes for space travel are moral, the longstanding ethical issues of single-use rockets and space junk still should have precautions and reforms. These two factors have contributed to the waste of Earth's resources along with the risks of endangering future spacecrafts that will have to out-maneuver the numerous floating metal objects in space. Past and current spaceflight is also associated with a great amount of carbon dioxide emissions that humanity has a responsibility to curb through multi-use rocket programs and spacecrafts that do not release carbon dioxide on lift off. If humanity decides to permanently follow through on the space industry for long-term success, we must not sacrifice Earth, its atmosphere, and future space missions from high amounts of space junk, all of which are degrading humanity's long-term potential.

Since its inception in the 1950s, the space industry has always been uber-expensive and meant to benefit largely and solely the nation that leads the project. In the modern day, this objective still pertains but, with new innovations, that "benefit" has been made clearer than ever before. As long as some of the precautions and ethical advice in this paper see resemblance in the real world, space deserves to have continued development as long as the long-term goal always circles back to serving humanity. We should not be living in a future where the government wastes billions of dollars on faulty programs, private companies develop space solely out of greed, Earth is needlessly over-polluted and over-extracted by rockets, and the edges of our planet is covered with space junk fated to stay in orbit for thousands of years. Thus, the rockets can rev their engines but with the seatbelt on, because the Space Race's Second Lap has practically just begun.

6. References

[1]National Air and Space Museum. "Missile, Surface-to-Surface, V-2 (A-4) | National Air and Space Museum." Airandspace.Si.Edu, 3 Aug. 2000, https://airandspace.si.edu/collection-objects/missile-surface-surface-v-2-4/nasm_A19600342000. Accessed 24 Aug. 2026.

[2]"What Was the Space Race?" National Air and Space Museum, Smithsonian, 23 Aug. 2023,

https://airandspace.si.edu/stories/editorial/what-was-space-race. Accessed 24 Aug. 2026.
[3]Dawsey, Jason. "Wernher von Braun and the Nazi Rocket Program: An Interview with Michael Neufeld, PhD, of the National Air and Space Museum." The National WWII Museum | New Orleans, 10 Sept. 2019, https://www.nationalww2museum.org/war/articles/wernher-von-braun-and-nazi-rocket-program-interview-michael-neufeld-phd-national-air. Accessed 24 Aug. 2026.
[4]"From WWII to the Space Race: The Story of Project Paperclip." Museum of Jewish Heritage — A Living Memorial to the Holocaust, 30 Aug. 2022, https://mjhnyc.org/events/from-wwii-to-the-space-race-the-story-of-project-paperclip/. Accessed 24 Aug. 2026.
[5]Space Center Houston. "Saturn V at Rocket Park." Space Center Houston, https://spacecenter.org/exhibits-and-experiences/nasa-tram-tour/saturn-v-at-rocket-park/. Accessed 24 Aug. 2026.
[6]NASA. "July 20, 1969: One Giant Leap for Mankind - NASA." *NASA*, NASA, 20 July 2019, https://www.nasa.gov/history/july-20-1969-one-giant-leap-for-mankind/. Accessed 11 Oct. 2026.
[7]National Air and Space Museum. "Why Did We Stop Going to the Moon?" Airandspace.Si.Edu, 9 Nov. 2021, https://airandspace.si.edu/stories/editorial/why-did-we-stop-going-moon. Accessed 24 Aug. 2026.
[8]NASA. "International Space Station - NASA." NASA, NASA, 23 May 2023, https://www.nasa.gov/reference/international-space-station/. Accessed 24 Aug. 2026.
[9]"Read." The National Academies Press, 2020, https://www.nationalacademies.org/read/18248/chapter/3#15. Accessed 25 Aug. 2026.
[10]"3 Things Driving the Commercial Space Boom." Johns Hopkins in Washington, D.C., 18 Nov. 2025, https://washingtondc.jhu.edu/news/drivers-of-the-space-economy/. Accessed 24 Aug. 2026.
[11]"The New Space Race." Aerospace America, 31 Mar. 2026, https://aerospaceamerica.aiaa.org/the-new-space-race/. Accessed 24 Aug. 2026.
[12]"The Space Report 2025 Q2 Highlights Record $613 Billion Global Space Economy for 2024, Driven by Strong Commercial Sector Growth." Space Foundation, 22 July 2025, https://www.spacefoundation.org/2025/07/22/the-space-report-2025-q2/. Accessed 24 Aug. 2026.
[13]Friends, Our. "Who Really Benefits When Private Space Companies Get Public Money." DCReport.Org, 12 May 2026, https://www.dcreport.org/2026/05/12/who-really-benefits-when-private-space-companies-get-public-money/. Accessed 24 Aug. 2026.
[14]Isakowitz, Steve. "Satellites: Driving A Burgeoning Space Economy." Aerospace America, 1 Mar. 2020, https://aerospaceamerica.aiaa.org/departments/satellites-driving-a-burgeoning-space-economy/. Accessed 24 Aug. 2026.
[15]"Artemis - NASA." Nasa.Gov, 2024, https://www.nasa.gov/humans-in-space/artemis/. Accessed 24 Aug. 2026.
[16]"Circular Economy to Tackle Space Junk | University of Southampton." Www.Southampton.Ac.Uk, https://www.southampton.ac.uk/news/2023/02/space-junk.page. Accessed 24 Aug. 2026.
[17]Dreier, Casey. "Your Guide to NASA's Budget." The Planetary Society, The Planetary Society, 2025, https://www.planetary.org/space-policy/nasa-budget. Accessed 24 Aug. 2026.
[18]Hamaker JW. The Faster, Better, Cheaper Approach to Space Missions: An Engineering Management Assessment. Nasa.gov1999. Accessed August 25, 2026. https://ntrs.nasa.gov/citations/19990100660
[19]Space Voyage Ventures Team. "Discover the Role of Private Companies in Space

Exploration." *Space Voyage Ventures*, 9 Jan. 2024, https://spacevoyageventures.com/role-of-private-companies-in-space-exploration/#examining-the-role-of-private-companies-in-space-exploration-0. Accessed 25 Aug. 2026

[20]Choi, Charles Q. "Rocket Launch Costs Have Dropped 96% since 1960. How Low Can They Go?" Space, 20 July 2026, https://www.space.com/space-exploration/launches-spacecraft/rocket-launch-costs-have-dropped-96-percent-since-1960-how-low-can-they-go. Accessed 24 Aug. 2026.

[21]"New Shepard Astronauts | Blue Origin." Blue Origin, 2021, https://www.blueorigin.com/new-shepard/astronauts. Accessed 24 Aug. 2026.

[22]"SpaceX Suffers Rare, Last-Second Falcon 9 Rocket Launch Abort – Spaceflight Now." Spaceflightnow.Com, 2026, https://spaceflightnow.com/2026/07/20/live-coverage-spacex-to-launch-24-starlink-satellites-on-falcon-9-rocket-from-vandenberg-sfb-11/. Accessed 25 Aug. 2026.

[23]Commercial Crew Program - NASA. https://www.nasa.gov/humans-in-space/commercial-space/commercial-crew-program/. Accessed 24 Aug. 2026.

[24]Paquin, Kelsey. "Private Sector Engagement." NASA Science, 10 June 2026, https://science.nasa.gov/earth-science/private-sector-engagement/. Accessed 24 Aug. 2026.

[25]"Boeing's Starliner Losses Reach $2 Billion." Spacepolicyonline.Com, 2025, https://spacepolicyonline.com/news/boeings-starliner-losses-reach-2-billion/. Accessed 24 Aug. 2026.

[26]Fed-Spend Intelligence. "SpaceX Government Contracts: $22 Billion in Federal Awards from NASA, DOD, and the Space Force." Fed-Spend.Com, 17 Mar. 2026, https://fed-spend.com/blog/spacex-government-contracts-nasa-dod-space-force. Accessed 24 Aug. 2026.

[27]NASA. "Benefits to Humanity - NASA." NASA, 14 Dec. 2023, https://www.nasa.gov/humans-in-space/benefits-to-humanity/. Accessed 24 Aug. 2026.

[28]The Daily Orbit. "The International Space Station Weighs 420 Tons and Stretches the Length of an American Football Field —." *Space Daily*, 26 June 2026, https://spacedaily.com/j-v-the-international-space-station-weighs-420-tons-and-stretches-the-length-of-an-american-football-field-making-it-the-most-expensive-object-humanity-has-ever-built-at-roughly-150-billion/. Accessed 26 Aug. 2026.

[29]Adler, Doug. "Why Did NASA Retire the Space Shuttle?" Astronomy Magazine, 12 Nov. 2020, https://www.astronomy.com/space-exploration/why-did-nasa-retire-the-space-shuttle/. Accessed 24 Aug. 2026.

[30]Byerly, Radford. "Space Shuttles Challenger and Columbia Accidents | Encyclopedia.Com." Encyclopedia.Com, 2019, https://www.encyclopedia.com/science/encyclopedias-almanacs-transcripts-and-maps/space-shuttles-challenger-and-columbia-accidents. Accessed 24 Aug. 2026.

[31]O'Rourke, Tim. "Chronicle Covers: When the Challenger Explosion Shook Us All." *San Francisco Chronicle*, SFGATE, 29 Jan. 2016, https://www.sfchronicle.com/chronicle_vault/article/Chronicle-Covers-When-the-Challenger-explosion-6791797.php. Accessed 11 Oct. 2026.

[32]SatBase. "SpaceX Increases Falcon 9 Launch Prices to $74M - Rideshare Now $7,000/Kg." SatBase, 28 Feb. 2026, https://satbase.com/articles/spacex-falcon-9-price-increase-2026. Accessed 24 Aug. 2026.

[33]"Statistics - SpaceXNow." Spacexnow.Com, 2024, https://spacexnow.com/stats. Accessed 24 Aug. 2026.

[34]"The Evolution of Space Travel: Then and Now | Richard Nixon Museum and Library." Www.Nixonlibrary.Gov, 5 Aug. 2021, https://www.nixonlibrary.gov/news/evolution-sp

ace-travel-then-and-now. Accessed 24 Aug. 2026.
[35]"Top Spaceflight Companies Pushing Human Exploration Forward." Celestis.Com, 2023, https://www.celestis.com/blog/top-spaceflight-companies/. Accessed 24 Aug. 2026.
[36]National Museum of the United States Air Force. "Cold War in Space: Top Secret Reconnaissance Satellites Revealed." National Museum of the United States Air ForceTM, https://www.nationalmuseum.af.mil/Visit/Museum-Exhibits/Fact-Sheets/Display/Article/195923/cold-war-in-space-top-secret-reconnaissance-satellites-revealed/. Accessed 24 Aug. 2026.
[37]Swope, Clayton. "The Future of Military Power Is Space Power." Csis.Org, 2025, https://www.csis.org/analysis/future-military-power-space-power. Accessed 25 Aug. 2026.
[38]Bohnert, Michael. "Cruise Missiles Are the Present and Future of Warfare." Rand.Org, Military Times, 9 Oct. 2025, https://www.rand.org/pubs/commentary/2025/10/cruise-missiles-are-the-present-and-future-of-warfare.html. Accessed 25 Aug. 2026.
[39]Keller, Jared. "The Pentagon's New Super Weapon Is Basically A Weaponized Meteor Strike." Task & Purpose, 7 June 2017, https://taskandpurpose.com/news/kinetic-bombardment-kep-weaponry/. Accessed 25 Aug. 2026.
[40]Pike, John. "Hypervelocity Rod Bundles (HRB)." *Globalsecurity.Org*, 2026, https://www.globalsecurity.org/space/systems/rods-from-god.htm. Accessed 11 Oct. 2026.
[41]Office, Accountability. "DOD's Space-Based Laser Program--Potential, Progress, and Problems." *Gao.Gov*, 2022, https://www.gao.gov/products/c-masad-82-10. Accessed 27 Oct. 2026.
[42]PBS News. "U.S. Companies Work to Cut Reliance on China for Critical Minerals Used in Weapons." PBS News, 3 Aug. 2026, https://www.pbs.org/newshour/nation/u-s-companies-work-to-cut-reliance-on-china-for-critical-minerals-used-in-weapons. Accessed 25 Aug. 2026.
[43]"Tungsten Deposits in the United States | U.S. Geological Survey." Www.Usgs.Gov, https://www.usgs.gov/data/tungsten-deposits-united-states. Accessed 25 Aug. 2026.
[44]America, Spirit of. "Spirit of America Survey: How Americans View America's Role in the World." Spirit of America, 3 Feb. 2025, https://spiritofamerica.org/how-americans-view-americas-role-in-the-world. Accessed 25 Aug. 2026.
[45]Observations and Science Potentially Enabled by the Vision for Space Exploration National Academies of Sciences, Engineering, and Medicine. 2007. The Scientific Context for Exploration of the Moon. Washington, DC: The National Academies Press. https://doi.org/10.17226/11954.
[46]Wattles, Jackie. "NASA Makes Moves to Dodge Costly Delays on Its Path to Build a $30 Billion Moon Base." CNN, 30 June 2026, https://www.cnn.com/2026/06/30/science/nasa-moon-base-announcement. Accessed 24 Aug. 2026.
[47]"Moon Base - NASA." NASA, 26 May 2026, https://www.nasa.gov/reference/moonbase-about/. Accessed 24 Aug. 2026.
[48]SpaceX. "SpaceX." SpaceX, 2025, https://www.spacex.com/humanspaceflight/moon. Accessed 24 Aug. 2026.
[49]Zubrin, Robert. "The Case for Colonizing Mars, by Robert Zubrin – NSS." National Space Society, Aug. 1996, https://nss.org/the-case-for-colonizing-mars-by-robert-zubrin/. Accessed 24 Aug. 2026.
[50]Cockell, Charles. "Mars Is an Awful Place to Live - NERC Open Research Archive." Nerc.Ac.Uk, 2024, https://doi.org/ Cockell, Charles. 2002 Mars is an awful place to live. Interdisciplinary Science Reviews, 27 (1). 32-38. https://doi.org/10.1179/030801802225002881

<https://doi.org/10.1179/03080180222500288 1>
[51]Cranford, Nathan, and Jennifer Turner. "The Human Body in Space." NASA, 2 Feb. 2021, https://www.nasa.gov/humans-in-space/the-human-body-in-space/. Accessed 24 Aug. 2026.
[52]"Mars Base by Dotted Yeti." *Humanmars.Net*, 11 Sept. 2024, https://www.humanmars.net/2020/04/mars-base-by-dotted-yeti.html. Accessed 11 Oct. 2026.
[53]Long, Jessica. "7 Surprising Facts about the Circular Economy for COP26." World Economic Forum, 27 Oct. 2021, https://www.weforum.org/stories/climate-action/7-surprising-facts-to-know-about-the-circular-economy-for-cop26/. Accessed 24 Aug. 2026.
[54]Record. "Record Number of Objects Launched Into Space Last Year." Yale E360, 2026, https://e360.yale.edu/digest/2025-satellite-launches. Accessed 24 Aug. 2026.
[55]West, Jessica, et al. "The Importance of Satellites to Life on Earth." Www.Cigionline.Org, 29 Jan. 2023, https://www.cigionline.org/multimedia/the-importance-of-satellites-to-life-on-earth/. Accessed 24 Aug. 2026.
[56]"Starlink." Starlink.Com, 2026, https://starlink.com/?srsltid=AfmBOoot3kSBV71uyHMhKzk0oHORdINtOJLTByic2nfMMNNnjc1m2yYC. Accessed 24 Aug. 2026.
[57]Margolis, Emily. "Space Tourism: Then and Now." *Airandspace.Si.Edu*, 25 Oct. 2021, https://airandspace.si.edu/stories/editorial/space-tourism-then-and-now. Accessed 25 Aug. 2026.
[58]"Charting Each Generation of NASA Astronauts Reveals How the 'Right Stuff' Has Changed." National Geographic, https://www.nationalgeographic.com/science/graphics/charting-how-nasa-astronaut-demographics-have-changed-over-time. Accessed 24 Aug. 2026.
[59]Hawley, Dustin. "Average Weight Of A Car." J.D. Power, 11 Dec. 2022, https://www.jdpower.com/cars/shopping-guides/average-weight-of-a-car. Accessed 24 Aug. 2026.
[60]"Critical Minerals in Space | SFA (Oxford)." SFA (Oxford), 2025, https://www.sfa-oxford.com/knowledge-and-insights/critical-minerals-in-low-carbon-and-future-technologies/critical-minerals-for-space-and-technology/. Accessed 25 Aug. 2026.
[61]Zisk, Rachael. "The Space Industry's Climate Impact: Part 2." Payload, 15 Dec. 2023, https://payloadspace.com/the-space-industrys-climate-impact-part-2/. Accessed 25 Aug. 2026.
[62]"A Guide to Metals Used in Spacecraft to Withstand High Temperature - Advantage Metal Products." Advantage Metal Products, 27 June 2024, https://advantagemetal.com/a-guide-to-metals-used-in-spacecraft-to-withstand-high-temperature/. Accessed 25 Aug. 2026.
[63]Analysis of Environmental Impact in Titanium Alloy vs Stainless Steel Production. Patsnp Eureka2025. Accessed August 25, 2026. https://eureka.patsnap.com/report-analysis-of-environmental-impact-in-titanium-alloy-vs-stainless-steel-production
[64]Ahmad, Shabbir, et al. "Carbon Emissions in Metal Manufacturing Productivity: A Global Analysis of Aluminium Smelting." Energy Research &Amp; Social Science, vol. 129, Elsevier BV, Nov. 2025, p. 104343, https://doi.org/10.1016/j.erss.2025.104343.
[65]NASA. "Space Nuclear Propulsion - NASA." NASA, 8 May 2023, https://www.nasa.gov/space-technology-mission-directorate/tdm/space-nuclear-propulsion/. Accessed 24 Aug. 2026.
[66]Muransky, Ondrej. "Nuclear Propulsion Systems | ANSTO." Www.Ansto.Gov.Au, 2023, https://www.ansto.gov.au/our-science/nuclear-technologies/reactor-systems/nuclear-propulsion-systems. Accessed 24 Aug. 2026.
[67]NASA. "4.0 In-Space Propulsion - NASA." NASA, 17 Mar. 2024,

https://www.nasa.gov/smallsat-institute/sst-soa/in-space_propulsion/. Accessed 24 Aug. 2026.
[68]Quach, Phu. "How Reusable Rockets Land And Why It Matters - USC Viterbi School of Engineering." USC Viterbi School of Engineering - USC Viterbi School of Engineering, 26 Feb. 2026, https://illumin.usc.edu/how-reusable-rockets-land-and-why-it-matters/. Accessed 24 Aug. 2026.
[69]Boex, Chris.The recycled space shuttle parts that will power Artemis II towards the moon. LiveNOW from FOX | Breaking News, Live EventsApril 2026. Accessed August 25, 2026. https://www.livenowfox.com/news/artemis-rocket-space-shuttle-engines
[70]STEM. It's Not Rocket Science, Or Is It? catalyst-magazine.org. Accessed August 25, 2026. https://catalyst-magazine.org/articles/its-not-rocket-science-or-is-it/
[71]1.O'Callaghan J. What is Space Junk and why is it a problem? Natural History Museum. Accessed August 25, 2026. https://www.nhm.ac.uk/discover/what-is-space-junk-and-why-is-it-a-problem.html
[72]Berkley, Ben. "There's Just so Much Space Junk." *Thehustle.Co*, HubSpot, 9 Oct. 2023, https://thehustle.co/there-s-just-so-much-space-junk. Accessed 11 Oct. 2026.

Journal of Ethics in Scientific and Technological Innovations

What It Takes to Sell: An Economic and Ethical Evaluation of Planned Obsolescence in Consumer Electronics

Melody Dai[1*], Janet Guan[2*]
[1]BASIS Independent Silicon Valley, San Jose, CA 95126, [2]Mission San Jose High School, Fremont, CA 94539

Abstract

Planned obsolescence (PO), the intentional shortening of a product's lifespan to spur replacement buying, is a controversial business strategy that has been employed in industries ranging from fashion to automobile production. In particular, PO has been increasingly used by the consumer electronics industry. This paper examines two main categories of PO: (I) technological obsolescence, or deliberately limiting product functionality, and (II) psychological obsolescence, or manipulating a product's perceived value. Arguments defending PO as necessary for economic regulation and innovation are explored.[1, 2] Additionally, the consumer and environmental impacts of PO are discussed; specifically, PO may hurt consumer finances, reduce producer transparency, and exploit consumer irrationality, while contributing to waste electrical and electronic equipment accumulation and pollution.[3, 4, 5] Finally, past legal responses to PO are addressed. Through evaluating cases of PO in popular consumer electronics, this paper lends to existing discussion on the motivations and ethical implications of PO.

1. Introduction

"Planned obsolescence" (PO) is the intentional shortening of a product's lifespan by its producer to stimulate consumer demand for a replacement product.[3, 1] Examples of PO can be found in various industries, including fashion and automobile production. With the rise in demand for consumer electronics in the past decade, however, PO has become increasingly relevant in electronics manufacturing and marketing.

There exist multiple methods for implementing PO, among which are limiting product durability, designing a product to be difficult to repair, and routinely releasing new products with marginal updates. Various papers have categorized PO methods in different ways; Maycroft (2009) separated PO into technological, stylistic, and superfluous obsolescence, or relating to technological function, aesthetic fashionability, and additional purchase requirements;[6] Malinauskaite and Erdem (2021) used technological and stylistic (or psychological) obsolescence as primary distinctions, with stylistic encompassing general fashionability beyond aesthetic appeal.[4] This paper defines and uses two broader categories of PO: (I) technological and (II) psychological obsolescence (Fig. 1). Technological obsolescence refers to all aspects of necessary product function, including durability, compatibility, and repairability. Psychological obsolescence alters a product's perceived value using fashionability and signaled newness.

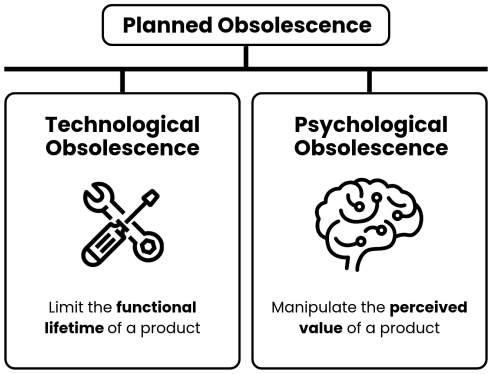


Fig 1. Types of planned obsolescence: (I) technological (II) psychological

PO has been defended as necessary against economic underconsumption and for innovation.[4] At the same time, PO has been criticized for burdening consumers financially and contributing to waste electrical and electronic equipment (WEEE) accumulation.[3, 5] This paper aims to analyze the impact of various PO methods using case studies on popular consumer electronics, as well as investigate the economic and ethical implications of the practice.

2. PO Methods

2.1 Technological Obsolescence

We use the term "technological obsolescence" to denote PO methods by which producers intentionally limit the functional lifetime of a product. This can be done through deliberately reducing the durability or designing a product for limited uses in order to encourage repeated consumption of a product. Technological obsolescence also encompasses incompatibility obsolescence, by which a product is rendered impractical to use or useless due to incompatibility with newer products or services, such as software updates.[7] In addition, designs which make repair difficult or impossible, such as through limited access to replacement parts, fall under technological obsolescence.[4]

Particularly for consumer electronics, software updates are commonly used to reduce the durability or performance of a product. In 2020, Apple admitted to reducing CPU performance in older iPhone models, despite previously having denied doing so, in order to preserve battery health after receiving customer complaints of unexpected device shutdowns.[8] The device slowdowns reportedly affected models released from 2014 to 2016, driving consumers to replace the sluggish devices by purchasing newer models which "potentially boosted sales by millions of devices per year." Similarly, in 2018, the Italian Competition Authority (ICA) fined Apple and Samsung for repeatedly asking consumers to upgrade firmware on the iPhone 6, 6Plus, 6s and 6sPlus and Galaxy Note 4 respectively, which resulted in reduced device performance that consumers had not been sufficiently informed about prior.[9] In both cases, the new firmware was optimized for newer models, which caused greater strain on older hardware that led to malfunctions and shutdowns. The ICA additionally found that Apple had failed to provide adequate assistance or information about restoring device function, such as through downgrading or battery substitution. Finally, Samsung was determined to have been intentionally refusing assistance to consumers whose products were no longer under warranty and charging high prices for repairs with no option to downgrade. Both cases caused consumers to replace devices by purchasing new phones, as the old devices had been rendered dysfunctional through software update induced technological obsolescence.

Conversely, software update restrictions are a method of incompatibility-driven technological PO. The iOS 16 update in 2022 did not support the iPhone 7, which had been discontinued a mere three years prior, and earlier devices, raising concerns about device security.[10] Similarly, Macbooks historically stop receiving macOS updates roughly five to seven years after release, with security updates continuing for another two to three years.[11] For example, updates for Macbook models from 2015 to 2017 were removed when the macOS Ventura was released in 2022.[10] Not only do consumers not get the latest features, but they also miss critical security updates which patch publicly known vulnerabilities that attackers exploit, hence many consumers choose to replace devices after software updates are no longer supported for their devices.

Reparability suppression as a technological PO method comes in various

forms, ranging from situations where a product cannot be repaired without destroying or impeding its functionality and where certain construction choices make repair extremely challenging or costly. In 2003, Apple themselves explained that the high price for the iPod's battery replacement program ($99) was because the product was designed to be cheaper to replace than to repair.[12] As a more modern example, AirPods are not only designed to only last 18 to 36 months but also are impossible to repair without cutting open and thus destroying the casing.[13] In 2021, consumers discovered that replacing the iPhone 13's screen completely disabled Face ID (Fig. 2), causing critical security concerns and creating a strong case for right to repair legislation.[14] While the issue was never officially addressed in the release notes, the iOS 15.2 update would later resolve the Face ID malfunctions due to repair. Over the years, Macbooks have also faced growing challenges to repairability, with motherboards being soldered for Macbook Air models from 2008 onwards and hard drives being soldered for Macbooks from 2015 onwards.[15] Being unable to repair their devices, consumers are forced to replace them by purchasing a new one.



Fig 2. Deactivation of Face ID following phone screen repair[14]

It should be acknowledged that in many cases, it may not be the intention or primary intention, such as when aesthetics or costs are prioritized, of the producer to render a product obsolete by reducing the durability, compatibility, or repairability of a product. There are situations where a product shows such significant improvement, whether by function or safety, that an entire category of products can become obsolete.[6] Regardless of planned or unplanned nature, however, technological obsolescence has consistently shown to induce consumers to purchase new products due to concerns over functionality and security.

2.2 Psychological Obsolescence

We use the term "psychological obsolescence" to refer to PO methods by which producers manipulate consumer perception of a product, such as its value or outdatedness, without necessarily altering its functionality. Psychological obsolescence is executed at the product post-release stage typically through intense marketing campaigns.[4] The underlying strategy behind this type of PO is to disguise producer intent as consumer needs; Maycroft (2009) described the tactics as "convinc[ing consumers] that regular stylistic changes were an opportunity rather than a costly inconvenience".[6] This type of PO is most closely associated with culture and social behaviors. Producers exploit the imperfect rationality of consumers and thus hold significant power over demand.[4]

One primary method of psychological obsolescence is fashionability of product features and aesthetic appeal. Fashionability relies on the tendency for consumers to view a product as necessary for social inclusion or as a status symbol.[6] General Motors (GM) overtook Ford in the 1920s, for example, as the largest American automobile manufacturer by simply releasing new colors and incrementally increasing the speed of their models annually.[16] GM created a standard of fashionability that motivated Americans to continue buying new models year after year. Consumer electronics

today continue to develop marketing techniques around identity and status consumption patterns. Apple releases new iPhone models annually with few major functional changes; Ribeiro et al. (2024) cited multiple tech and news sources that described the modifications in iPhones 12, 13, and 14 as "iterative updates" compared to iPhone 11.[7]

Producers can also use psychological obsolescence to hide accompanying technological obsolescence.[6] The iPhone 13 screen repair issue discussed in 2.1 occurred as Apple continued to make iterative and stylistic updates to their phone line.[16] The adoption of nonremovable batteries in smartphones is another example; the battery is difficult to replace, but designs can be 25% thinner for aesthetic appeal.[17] Connected to this is the prioritization of form over function in product design, which can be motivated by the stylistic aspect of psychological obsolescence.[6] For instance, the Apple Magic Mouse 2 has a sleek, glassy surface but can only be charged from its bottom, rendering it unusable while charging.[18] Samsung's Frame TV, skinny and minimalistic in design to mimic an art display, has received user reports of frequent glitches.[19]

Interestingly, there is an element of psychological obsolescence within all methods of PO in the form of perceived outdatedness. Kupperweiser et al. (2019) found that the announcement of a new product alone, indicating potential outdatedness, reduces the value perceived by consumers of the current product.[3] In other words, signaled outdatedness without any changes to the current product can already alter consumer behavior.

Overall, psychological obsolescence encompasses forms of PO that involve consumer choice but still ultimately lend control to producers.

3. Economic Defenses

3.1 Economic Underconsumption

PO has been argued to be a necessary component of a sustainable economy. Moyse (2021) described PO as a "natural companion to the market economy," where high product replacement rates can signal a healthy economy.[1] However, when PO was first coined by American real estate agent Bernard London during the Great Depression in his 1932 paper "Ending the Depression Through Planned Obsolescence," his principal concern was underconsumption.[20]

Multiple high-profile economists attributed the Great Depression to low expenditures. John Maynard Keynes specifically argued so in his 1936 book *The General Theory of Employment, Interest and Money.*[21] London's paper posited the legal implementation of PO, or establishing laws to explicitly restrict the economic lifespans of products, as a mechanism for reinvigorating the depressed economy by necessitating increased consumer spending and production, the latter consequence also boosting employment rates.[20]

London's theory of PO coincided with other ideas for economic regulation at the time, including that of economist Thorstein Veblen. Veblen had established his theory of the "Soviet of Technicians," or placing engineers and scientists in government and entrepreneurial roles, a response to his view that the corrupt practices of profit-oriented businessmen were responsible for the Great Depression.[1] Veblen argued that engineers, who understood the technical aspects of products being exchanged and thus their social implications, could instill order within the economy and be trusted to abide by social responsibility.[22] Moyse (2021) identifies similarities between Veblen's ideas and London's PO in spirit, where the opportunity to design for PO would give engineers greater control of the economy.[1]

Notably, Moyse (2021) highlighted the major assumption in London's theory that corporations would necessarily prioritize social responsibility and welfare.[1] While economic stability remains a concern today, there is a difference in context between when PO was first defined and now. Nearly a century after the Great Depression and with the widespread adoption of PO as a business strategy by large corporations, it is pertinent to weigh the proposed economic benefits of PO against the theoretical risk for underconsumption with the existing ethical consequences of PO overuse.

3.2 Innovation

Another prominent defense of PO is its role in innovation and technological progress. Howell (1962) described PO as "an undeniable consequence" of planned innovation, or the "management of technological change".[23] Fishman et. al (1993) demonstrated in their economic model that PO increases the incentive for producers to invest in innovation, where without PO, consumers are only willing to pay for innovative benefits and not additional costs of production because their old products are still durable and thus continue to provide utility.[24] Fishman et al. (1993) emphasized the case where an innovative change is significant to social welfare but not profitable enough to outweigh production costs, hence deterring producers from making the change.

While the underlying regulating effect of PO on product economic lifespans may sustain innovation, not all instances of PO offer significant innovative benefits. Iterative product updates, as discussed in 2.2 in regards to Apple and Samsung smartphones, make only marginal changes to new products.[7] Additionally, conflicts about intellectual property associated with PO, such as right to repair as discussed in Section 2.1, may reinforce competitive barriers and restrict collaborative efforts to advance innovation. PO should therefore be evaluated through both the lens of overall social responsibility as well as individual producer intent.

4. Ethical Implications

4.1 Consumer Impact

The connection between consumerist culture and the disposability intrinsic to regular replacement buying may suggest that consumer behavior rewards PO. Ludwig von Mises (2007) proposed that consumers prefer shorter product replacement intervals because they offer more immediate satisfaction.[25] This perspective raises the question of whether consumer satisfaction in the short term justifies long-term financial detriment. PO functions by increasing the frequency of replacement buying, which may lead consumers to make otherwise unnecessary purchases. Certain cases of technological obsolescence can further increase these added expenditures; producers that monopolize repair services may charge higher prices, and incompatibility may force consumers to buy additional products that are compatible. These costs can accumulate over time and most drastically impact consumers from lower income brackets, who may not be able to afford frequent product replacements.[4]

Various PO techniques are also inherently manipulative. First, technological obsolescence leads to poor transparency between companies and consumers. In the 2020 Apple battery slowdown case discussed in 2.1, state investigators in Arizona found that the company deliberately hid the slowdowns before later admitting responsibility, which impeded device repair and forced consumers to purchase replacements.[8] In the 2018 ICA cases discussed in 2.1 consumers reported poor product support, with Samsung allegedly refusing to provide support altogether.[9] This lack of transparency can be especially concerning in regards to security functions. When iOS 15 was released in 2021, Apple told consumers that iOS 14 would

"still get important security updates for a period of time".[10] The same announcement remained on Apple's iOS feature website page for at least a year, even though the company stopped providing updates after 36 days. 17% of iPhone users in May 2022 were still using iOS 14, likely unaware that the updates were paused and their devices were vulnerable. It should be noted that the discontinuation of updates may not pose an issue for the majority of consumers, as in a study by Adrion and Woidasky (2018), the mean usage length for portable devices, such as laptops and notebooks, was 4.0 years.[12]

Second, psychological obsolescence capitalizes on consumer irrationality. Malinauskaite and Erdem (2021) compares the deliberate influence producers impose upon consumer needs to a "confusopoly," where producers use intentional obfuscation to confuse consumers into making suboptimal purchasing decisions.[4, 26] Becher and Sibony (2021) describes the bandwagon effect as another tool within psychological obsolescence, where "people to behave in a certain way … mainly because others are behaving that way … reflecting herd behavior".[27] Psychological obsolescence may also employ marketing techniques that reinforce standards of fashionability, contributing to consumerist culture that tempts consumers into making unwise financial decisions.

While PO relies on consumer choice, because actors of PO have the capacity to greatly influence consumer needs and decisions, producers should be cognizant of the impact PO techniques have on consumer wellbeing. Necessary transparency should be given when consumer safety is at risk, and marketing techniques should take into account existing consumer needs before artificially imposing new ones.

4.2 Environmental and Health Impacts

Aside from the economic and social effects of PO, the waste electrical and electronic equipment (WEEE) generated also poses significant environmental and health concerns. Around 50 million tons of WEEE was generated in 2019, with this number projected to increase to 111 million tons by 2050.[5] By comparison, only 8.9 million tons of WEEE are properly recycled each year.[28]

In fact, WEEE is among the fastest growing types of waste not only in size but also toxicity, releasing toxic chemicals including heavy metals like cadmium and lead that mobile phones have shown an increase in from 2001 to 2015.[29, 30] When improperly recycled, such toxic substances not only create environmental pollution, but are also a serious health hazard, causing digestive, respiratory, and cardiovascular illnesses.[31] Furthermore, recycling of WEEE was shown to have a clear environmental advantage compared to a baseline of total incineration of WEEE when examining Swiss take-back and recycling programs for WEEE.[32]

Because WEEE is difficult to properly process, it is often shipped to developing countries, where improper WEEE management exacerbates environmental and health hazards. 23% of WEEE generated in developed countries—with the US, EU 25 countries, and Japan being leading producers and exporters of WEEE—is transported to developing regions in Africa and Asia, namely Ghana, Nigeria, South Africa, Vietnam, India, Philippines, and China.[33, 34] Many e-waste workers in these countries, many of whom are minors, are unaware of the level of toxicity of the materials they are exposed to, working without any or with minimal protective equipment.[34] Toxins contaminate both nearby residents and the environment, including the water and agriculture; studies from e-waste processing sites such as Guiyu, China have found people in

these areas to have unacceptable levels of toxins in their tissue and significantly greater exposure to pollutants than than those from control areas.[35] Certain methods of WEEE management are more damaging than others, with landfilling and open burning being amongst the least favorable practices.[36] For these reasons, the Basel Convention and similar legislation place restrictions on the transport of WEEE with the goal of keeping WEEE in producer countries or countries with the capacity to safely process it.[34, 35]

Efforts have been made to prevent the hazards of WEEE by lowering its toxicity, which decreases its health and environmental risks while promoting recyclability. The Restriction of Hazardous Substances Directive (RoHS) was enacted by the EU to restrict the amount of cadmium, mercury lead, hexavalent chromium, polybrominated biphenyls (PBB), polybrominated diphenyl ethers (PBDE), bis(2-ethylhexyl) phthalate (DEHP), butyl benzyl phthalate (BBP), dibutyl phthalate (DBP), and diisobutyl phthalate (DIBP) used in electrical and electronic equipment.[37] Many other countries, including China and India, have adapted similar directives, all of which work in tandem with take-back and recycling systems, such as those established by the EU WEEE Directive.[34]

Institutions and consumers can also help mitigate the effects of PO by reducing the amount of WEEE generated in the first place. Implementing a Circular Economy in the WEEE industry would support more reusing, repairing, and recycling of materials and products in order to increase the life-cycle of products, simultaneously reducing waste.[4, 38] Significant research and planning has been done on this topic already, with the European Commission (EC) preparing to revise the WEEE Directive in the upcoming Circular Economies Act.[39] The responsibility also lies in part with the consumer, as the European Environmental Agency reported that while smartphone designed life had increased the past few years, the designed lifetime was rarely reached due to wear and tear from use and due to consumers wanting the newest models.[40] Cooperation and informed decision-making and response at all levels is crucial to reduce the effects of PO.[41]

4.3 Legal Responses

Governments have combined prevention and reaction to restrict PO in the consumer electronics industry, as well as mitigate the practice's effect on consumers and the environment. Prevention has focused on the design stage. The EU's Ecodesign Directive outlines ecodesign for device longevity to limit waste and promote recyclability.[4] The EC's WEEE Directive standardized portable chargers to USB Type-C port in 2022,[42] curbing PO that relies on device incompatibility. Governments have also considered durability requirements to address general PO; the EU's Ecodesign Directive has set durability requirements for lighting and vacuum cleaners.[43] Maitre-Ekern et al. (2016) has noted, however, that such requirements are often difficult to concretely define because durability involves multiple different metrics. For products undergoing technological development, these standards also have to be adjusted for new models. Additionally, an organized system of oversight is needed to enforce the requirements.

Reaction has centered on consumer rights and support. France has targeted transparency by requiring producers to list information regarding the availability of spare parts on device labels if given by manufacturers.[43] Likewise, the EC has stated that it will "specifically consider . . . durability information in future Energy Labelling measures." It is important to note that labeling alone does not guarantee complete transparency; producers can fabricate label information without legislation that holds them directly

accountable. Another attempted approach is extending device warranties. Most of the European countries have warranties of two years, but Norway and Iceland impose five-year warranties for devices with longer expected lifespans; a 2007 ruling by the Norwegian Supreme Court found that smartphones with an expected lifespan of three to four years qualified for a five-year warranty.[43] Legislation has also strengthened consumer access to repair. France requires producers to provide spare parts listed as available on device labels to consumers within two months if needed for repair.[43] Sweden has reduced labor taxes for smartphone and computer repair.[4]

France is the first country to directly address PO legally; its Consumer Code defines PO, lists examples of PO methods, and fines producers using the practice 300,000 Euros.[43] In practice, it can be difficult for consumers to prove obsolescence to be deliberate, especially when indirect methods such as psychological obsolescence are used.

These approaches are a promising start to preventing PO overuse in the consumer electronics industry, promoting consumer rights, and addressing environmental consequences. Future legislation should focus on producer accountability, specifically to ensure sufficient oversight for enforcing durability and transparency requirements, and clarifying the distinction between PO and unintentional obsolescence.

5. Conclusion

The goal of this paper was to categorize and investigate PO use in the consumer electronics industry and its economic motivations and associated ethical issues.

Case studies of multiple products, such as smartphones, laptops, and related accessories, showed instances of both technological and psychological obsolescence in consumer electronics design. In particular, methods of technological obsolescence generally fell under issues with durability, compatibility, or repairability of devices, which psychological obsolescence was concerned with the fashionability and perceived outdatedness of a product.

Evaluation of PO defenses revealed the roles that PO may have in economic regulation and driving innovation, while also posing limitations in those defenses pertaining to shifts in economic and commercial situations over time as well as assumptions made about the behavior of producers and consumers. Ethical concerns about PO's impact on consumer choice and WEEE generation were also discussed along with past legal responses to these issues.

Overall, the use of PO in consumer electronics design has consistently been evidenced to drive consumers to regularly purchase and replace old devices, a tendency which can both create some economic benefits but also raise ethical concerns regarding consumers and the effects of electronic waste.

6. References

[1]Moyse, Pierre-Emmanuel. "The uneasy case of programmed obsolescence." *SSRN Electronic Journal*, 2020, https://doi.org/10.2139/ssrn.3804451.

[2]Schumpeter, Joseph A. *Capitalism, Socialism, and Democracy*. 3rd ed., Harper & Row, 1950.

[3]Kuppelwieser, Volker G., et al. "Consumer responses to planned obsolescence." *Journal of Retailing and Consumer Services*, vol. 47, Mar. 2019, pp. 157–165, https://doi.org/10.1016/j.jretconser.2018.11.014.

[4]Malinauskaite, Jurgita, and Fatih Buğra Erdem. "Planned obsolescence in the context of a holistic legal sphere and the circular economy." *Oxford Journal of Legal Studies*, vol. 41, no. 3, 7 Feb. 2021, pp. 719–749, https://doi.org/10.1093/ojls/gqaa061.

[5]Forti, Vanessa, et al. *The Global E-waste Monitor 2020*. United Nations University,

International Telecommunication Union, and International Solid Waste Association, 2020.
[6]Maycroft, Neil. *Consumption, Planned Obsolescence and Waste*. 2009.
[7]Ribeiro, Jorge, et al. "Apple and its management of innovation approach – to be or not to be obsolete?" *Communications in Computer and Information Science*, 2025, pp. 167–184, https://doi.org/10.1007/978-3-031-84078-4_13.
[8]Allyn, Bobby. "Apple Agrees to Pay $113 Million to Settle 'Batterygate' Case over iPhone Slowdowns." *NPR*, NPR, 18 Nov. 2020, www.npr.org/2020/11/18/936268845/apple-agrees-to-pay-113-million-to-settle-batterygate-case-over-iphone-slowdowns.
[9]De Franceschi, Alberto. "Planned obsolescence challenging the effectiveness of consumer law and the achievement of a sustainable economy." *Journal of European Consumer and Market Law*, vol. 7, no. 6, 2018, p. 217.
[10]McElhearn, Kirk, and Joshua Long. "Apple's Planned Obsolescence: iOS 16, macOS Ventura Drop Support for Many Models." *The Mac Security Blog*, 14 June 2022, https://www.intego.com/mac-security-blog/apples-planned-obsolescence/.
[11]Saini, Rajat. "How Long Your M1 Mac Really Lasts Before Apple Drops Support." *The Mac Observer*, 6 Dec. 2025, https://www.macobserver.com/news/how-long-your-m1-mac-really-lasts-before-apple-drops-support/.
[12]Zallio, Matteo, and Damon Berry. "Design and planned obsolescence. theories and approaches for designing enabling technologies." *The Design Journal*, vol. 20, no. sup1, 28 July 2017, https://doi.org/10.1080/14606925.2017.1352879
[13]Taffel, Sy. "Airpods and the Earth: Digital Technologies, planned obsolescence and the Capitalocene." *Environment and Planning E: Nature and Space*, vol. 6, no. 1, 31 Jan. 2022, pp. 433–454, https://doi.org/10.1177/25148486221076136.
[14]Purdy, Kevin. "Apple's New Screen Repair Trap Could Change the Repair Industry Forever." *iFixit*, iFixit, 27 Aug. 2026, www.ifixit.com/News/54829/apples-new-screen-repair-trap-could-change-the-repair-industry-forever.
[15]"Mac 101: How to Upgrade a Computer with Soldered Components." *Rocket Yard*, 16 May 2018, eshop.macsales.com/blog/44610-mac-101-how-to-upgrade-a-computer-with-soldered-components/?srsltid=AfmBOorMZH7HwWfuPWMmcZ5RMr1q7x-mmcpnVx8KyH8oWCmOPLJahZMF.
[16]"GM's Role in American Life." *NPR*, NPR, 2 Apr. 2009, www.npr.org/2009/04/02/102670076/gms-role-in-american-life.
[17]Adrion, Martin, and Jörg Woidasky. "Planned obsolescence in portable computers - empirical research results -." *Cascade Use in Technologies 2018*, 15 Aug. 2018, pp. 13–20, https://doi.org/10.1007/978-3-662-57886-5_3.
[18]Chambers-Ostler, Eliot. "Form over Function: Beautifully Designed Things That Simply Don't Work." *Heart Internet*, 30 Apr. 2019, www.heartinternet.uk/blog/form-over-function-beautifully-designed-things-that-simply-dont-work/.
[19]"An Honest Samsung Frame TV Review." *PhD in Clothes*, 8 Apr. 2026, phdinclothes.com/2025/08/02/samsung-frame-tv-review/.
[20]London, Bernard Miall. "Ending the Depression through Planned Obsolescence." *Revue du MAUSS*, no. 43, 2014, pp. 47-50.
[21]Keynes, John Maynard. *The General Theory of Employment Interest and Money*. Macmillan for the Royal Economic Society, 1973.
[22]Veblen, Thorstein. *The Vested Interests and the Common Man; and, the Engineers and the Price System*. Routledge/Thoemmes Press, 1994.
[23]Howell, Leland L. "Planned Innovation as a marketing tool." *California Management*

Review, vol. 4, no. 4, July 1962, pp. 66–71, https://doi.org/10.2307/41165500.
[24]Fishman, Arthur, et al. "Planned obsolescence as an engine of Technological Progress." *The Journal of Industrial Economics*, vol. 41, no. 4, Dec. 1993, p. 361, https://doi.org/10.2307/2950597.
[25]Mises, Ludwig von. Ludzkie działanie. Instytut Ludwiga von Misesa, 2007.
[26]Siciliani, Paolo, et al. *Consumer Theories of Harm*, 2019, https://doi.org/10.5040/9781509916887.
[27]Becher, Shmuel I., and Anne-Lise Sibony. *Confronting Product Obsolescence*, 2021, https://doi.org/10.2139/ssrn.3756452.
[28]Baldé, Cornelis P., et al. *The Global E-waste Monitor 2017*. United Nations University, International Telecommunication Union, and International Solid Waste Association, 2017.
[29]Gao, Ya, et al. "Global trends and future prospects of e-waste research: A Bibliometric analysis." *Environmental Science and Pollution Research*, vol. 26, no. 17, 29 Apr. 2019, pp. 17809–17820, https://doi.org/10.1007/s11356-019-05071-8.
[30]Singh, Narendra, et al. "Toxicity trends in e-waste: A comparative analysis of metals in discarded mobile phones." *Journal of Hazardous Materials*, vol. 380, Dec. 2019, p. 120898, https://doi.org/10.1016/j.jhazmat.2019.120898.
[31]Lin, Siyi, et al. "Toxic chemicals from uncontrolled e-waste recycling: Exposure, body burden, health impact." *Journal of Hazardous Materials*, vol. 426, Mar. 2022, p. 127792, https://doi.org/10.1016/j.jhazmat.2021.127792.
[32]Hischier, R., et al. "Does WEEE Recycling Make Sense from an Environmental Perspective?" *Environmental Impact Assessment Review*, vol. 25, no. 5, Jul. 2005, pp. 525–39. Crossref, https://doi.org/10.1016/j.eiar.2005.04.003.
[33]Vaccari, Mentore, et al. "WEEE Treatment in Developing Countries: Environmental Pollution and Health Consequences—An Overview." *International Journal of Environmental Research and Public Health*, vol. 16, no. 9, May 2019, p. 1595. Crossref, https://doi.org/10.3390/ijerph16091595.
[34]Sthiannopkao, Suthipong, and Ming Hung Wong. "Handling E-Waste in Developed and Developing Countries: Initiatives, Practices, and Consequences." *Science of The Total Environment*, vols. 463–464, Oct. 2013, pp. 1147–53. Crossref, https://doi.org/10.1016/j.scitotenv.2012.06.088.
[35]Adeola, Francis O. "WEEE Generation and the Consequences of Its Improper Disposal." *Waste Electrical and Electronic Equipment Recycling*, Elsevier, 2018, pp. 13–31. Crossref, https://doi.org/10.1016/B978-0-08-102057-9.00002-0.
[36]Singh, Narendra, et al. "Toxicity Evaluation of E-Waste Plastics and Potential Repercussions for Human Health." *Environment International*, vol. 137, Apr. 2020, p. 105559. Crossref, https://doi.org/10.1016/j.envint.2020.105559.
[37]European Commission. "RoHS Directive" *Environment - European Commission*, 2023. environment.ec.europa.eu/topics/waste-and-recycling/rohs-directive_en.
[38]Bressanelli, Gianmarco, et al. "Circular economy in the WEEE industry: A systematic literature review and a research agenda." *Sustainable Production and Consumption*, vol. 23, July 2020, pp. 174–188, https://doi.org/10.1016/j.spc.2020.05.007.
[39]European Commission. "Circular Economy." *Environment - European Commission*, 2025. environment.ec.europa.eu/strategy/circular-economy_en.
[40]Bachér, John, et al. *Electronic products and obsolescence in a circular economy*. European Topic Centre Waste and Materials in a Green Economy, 2020.
[41]Rivera, Julio L., and Amrine Lallmahomed. "Environmental implications of planned obsolescence and Product lifetime: A literature review." *International Journal of Sustainable*

Engineering, vol. 9, no. 2, 28 Dec. 2015, pp. 119–129, https://doi.org/10.1080/19397038.2015.109975.
[42]"The EU Common Charger." *Internal Market, Industry, Entrepreneurship and SMEs*, single-market-economy.ec.europa.eu/sectors/electrical-and-electronic-engineering-industries-eei/radio-equipment-directive-red/one-common-charging-solution-all_en.
[43]Maitre‐Ekern, Eléonore, and Carl Dalhammar. "Regulating planned obsolescence: A review of legal approaches to increase product durability and reparability in Europe." *Review of European, Comparative & International Environmental Law*, vol. 25, no. 3, Nov. 2016, pp. 378–394, https://doi.org/10.1111/reel.12182.

Digging the Hidden Toxicity: The Ethical Analysis of PFAS-Generated Risks in Daily Consumer Goods and Drinking Water

Megan Li[1*]
[1]*Zhengzhou Middle School, Zhengzhou, Henan, China*

Abstract

Per- and polyfluoroalkyl substances (PFAS) are persistent, bioaccumulative synthetic chemicals widely incorporated into daily consumer goods and industrial sectors, causing global drinking water contamination and long-term public health threats through exposure to daily-use products. Despite growing awareness of PFAS toxicity, global regulatory frameworks remain highly inconsistent, generating severe spatial environmental inequality between high-income nations and resource-limited regions. Furthermore, the lack of clear disclosure of added PFAS contents, insufficient non-target analysis, and the lack of cost-effective detection and purification methods together raise more ethical concerns. By evaluating the various influencing factors, this study reveals the main ethical debates between commercial production, corporate justice and transparency, technical requirements, human health, environmental sustainability, and global equity. By analyzing the ethical conflicts and concerns, this study also proposes several potential solutions and actions that could help improve human and environmental health while also promoting public equity worldwide.

1. Introduction

Per- and polyfluoroalkyl substances (PFAS) are a large group of fully synthetic fluorinated compounds that have been widely manufactured for global commercial and industrial use since the 1950s. [1,2] Consisting of a class of over 12,000 chemicals, PFAS include older long-chain chemicals such as Perfluorooctanoic acid (PFOA) and Perfluorooctanesulfonic acid (PFOS), and also some newer chemicals made of short-chain bonds such as hexafluoropropylene oxide dimer acid (commonly known as GenX, HFPO-DA).

Due to the existence of ultra-stable carbon–fluorine (C–F) covalent bonds, PFAS exhibit unique physical and chemical properties including exceptional thermal stability, water and oil repellency, high surface activity, and strong chemical inertness.[3] These properties make PFAS nearly ubiquitous in the environment and our daily life, as they are widely and efficiently used in industrial sectors and consumer goods (Fig. 1). In fact, PFAS are not only found in soil and lakes near factories but also in daily cookware such as non-stick cookware, various cosmetics, and even some waterproof outdoor jackets due to their strong resistance to water and oil.[4]

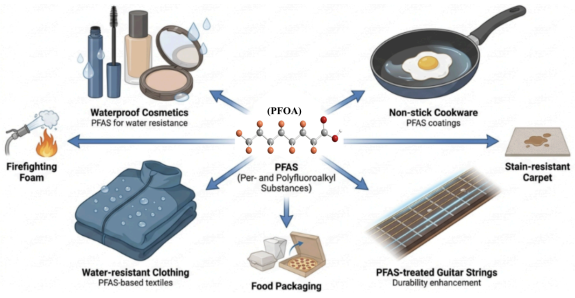


Figure 1: Illustration of the broad applications of PFAS in daily consumer goods.

However, also known as "forever chemicals", PFAS contain chains of carbon and

fluorine atoms which together form exceptionally strong C-F chemical bonds, making them extremely resistant to degradation and difficult to break down in both the environment and the human body.[5] Instead, they persist in soil, groundwater, and ecosystems for centuries. They also have high bioaccumulation potential and continuously accumulate in biota through the food web.[6]

Recently, an increasing number of studies have shown the potential risks of PFAS on human health and the entire ecosystem. PFAS can be detected in approximately ninety-seven percent of the U.S. population,[7] and they are also found globally at relatively high levels, raising growing health concerns among generations. While industrial applications such as firefighting foams and industrial coatings are well-documented sources of PFAS pollution, modern human exposure primarily stems from everyday contact with consumer products. Ordinary household and cosmetic products, including non-stick cookware, waterproof long-wearing makeup, and grease-resistant food packaging, routinely incorporate PFAS to enhance product durability and functionality.[2] Additionally, widespread industrial runoff contaminates municipal groundwater and public drinking-water systems, leading to chronic low-dose PFAS exposure for both animals and humans. Unlike acute environmental toxins that trigger immediate harm, PFAS pose prolonged hidden risks due to their strong bioaccumulative and biomagnifying nature; these compounds gradually accumulate in human blood, tissues, and food chains including livestock and dairy products,[8] which may lead to long-term immune dysfunction, developmental abnormalities,[9] type 2 diabetes,[10] and potential carcinogenic risks.[11]

Existing literature has mostly focused on examining the toxicological effects of PFAS, industrial pollution remediation, and technical detection optimization. As a result, it could be easy to overlook the "invisible" exposure to PFAS from daily consumer goods and drinking water. This study focuses instead on the ethical dimension of PFAS by analyzing the current status, existing problems, and ethical conflicts surrounding PFAS in consumer products and drinking water.

The ethical conflicts mostly lie between commercial production, corporation justice and transparency, technical requirements, human health, environmental sustainability, and global equity.

2. PFAS in Consumer Products

PFAS are currently widely used in daily consumer products such as waterproof long-wearing cosmetics, non-stick cooking utensils, and certain types of waterproof or grease-resistent outdoor clothes. This chapter evaluates the ethical sides of the effects of adding PFAS to these products, including health concerns, regulation inconsistency, global inequity, and the lack of clear content disclosure and comprehensive detection.

2.1 Two-sided Ethical Debates: technical convenience Vs. long-term health

2.1.1 Why are PFAS crucial to be added in many consumer products?

Consumer product manufacturers continue incorporating PFAS into daily goods primarily due to their unmatched and irreplaceable physicochemical properties which no low-cost alternative can currently fully replicate. Noted in the introduction, the extremely strong carbon–fluorine bonds give PFAS exceptional thermal stability, chemical inertness, and ultra-low surface tension, enabling unique hydrophobic and oleophobic characteristics that repel both water and oil.[3] These properties help PFAS perform well in products such as cosmetics and food packaging. In cosmetics, PFAS enhance the water-proof, sweat-proof, and long-wearing effects of

foundations, eyeliners, and lipsticks, significantly improving product durability and user experience.[12] In food packaging, PFAS greaseproof coatings prevent oil and moisture penetration, protecting packaged foods and extending shelf life.[13] Moreover, PFAS are even found in guitar strings, piano keys, and ski wax products.[12]

Due to the health risks and public concerns of PFAS, some companies have been developing alternative additives and production methods. However, it can be costly and time-consuming when researching and developing entirely new strategies, and the fluorine-free alternatives currently known are relatively expensive. In contrast, PFAS-based formulas are much more cost-effective, mature, and mass-producible. In other words, PFAS usage maximizes product marketability while minimizing manufacturing expenses, making it more advantageous for corporations.

2.1.2 Potential Health Risks

As mentioned in the Introduction and *2.1.1*, adding PFAS in daily products may be associated with a lot of health issues and potential risks including immune dysfunction, cancers, etc. Additional research has shown that dermal uptake is a main pathway for PFAS exposure, which further proves the side effects of adding PFAS in daily products. When using waterproof foundations, PFAS may penetrate the skin; short-chain PFAS show notable dermal permeability, entering the bloodstream after topical application.[14] For lip products, incidental oral ingestion during eating and drinking adds another exposure route.[15] In outdoor apparel, school uniforms and sportswear finished with durable water-repellent (DWR) PFAS coatings, chemicals migrate onto skin surface through friction and sweat, while volatile PFAS fragments can also be inhaled through fabric outgassing.[16,17] Though single-event exposure levels are often low, repeated daily contact produces chronic cumulative uptake, and PFAS bioaccumulates inside human tissues with slow elimination rates.[1] Children and infants are especially vulnerable groups, as their less-mature skin barrier increases dermal absorption risk, and hand-to-mouth behaviors raise additional oral-ingestion hazards from textile dust residues.[9,18] Whether to keep using trace amounts of PFAS for daily convenience or to pursue long-term health for future generations lies as the core conflict.

2.2 Global Inconsistency and Spatial Inequality

In recent years, many countries have introduced regulations setting minimum limits of added PFAS in consumer goods. However, there is a clear global inconsistency across developed countries and developing countries, which further reveals ethical issues.

The European Union (EU) has adopted ambitious class-oriented restrictions under Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) and Packaging and Packaging Waste Regulation (PPWR), setting strict concentration limits for PFAS in waterproof apparel, cosmetics, and grease-proof food packaging, and proposing broad restrictions covering thousands of PFAS variants except narrowly defined essential-use cases.[19] It has set a limit of 25 ppb for any individual non-polymeric PFAS, 250 ppb for the sum of non-polymeric PFAS, and 50 ppm for total PFAS, including polymeric PFAS. Besides, several individual EU member states implement even stricter national rules ahead of union-wide deadlines, creating minor divergences within Europe itself.[20] France has banned the use of PFAS in cosmetics since January 1st, 2026. In the United States, comprehensive federal restrictions on PFAS in everyday consumer products are absent; instead, rules vary by state, with some states banning intentionally added PFAS in cosmetics and textiles while other states maintain few enforceable limits.[15]

Nevertheless, many low- and middle-income countries lack dedicated legislation on PFAS in consumer products, enforcing partial bans for a small handful of legacy long-chain compounds such as PFOA and PFOS under the Stockholm Convention, with no systematic oversight for emerging short-chain PFAS substitutes.[21] In this case, trades in PFAS-containing products occur between developed and developing countries, meaning that PFAS hotspots may be shifting to low- and middle-income countries and cause severe health problems there.[22] Hazardous chemicals including PFAS are also illegally traded towards poorer areas with weak legislation; assessments show that about 27.5 million tonnes of banned hazardous chemicals were illegally transported to low- and middle- income countries between 2004 and 2019.[21]

This global regulatory inconsistency has severe environmental-justice consequences. Factories may continue producing PFAS-contained products to maintain the economic benefits, while selling those products with banned amounts of PFAS in some developing nations. This causes more health problems and environmental pollution, deepening spatial inequality and injustice in global public health. Additionally, since PFAS are persistent environmental hazards, products released in loosely-regulated regions still contribute to worldwide environmental circulation and cross-border contamination, meaning restrictive rules in wealthy regions cannot fully insulate global populations from PFAS harm.[1]

2.3 Lack of Clear Contents

According to previous investigations, most cosmetic products sold in the U.S. and in Canada, for instance, do not clearly label the presence of fluorinated ingredients.[23] Many customers buy and use these products without knowing that they may contain potentially hazardous additives. The lack of PFAS labels on consumer products is a crucial contributor to the spread of PFAS-related risks today.

2.4 Lack of Non-target PFAS Analysis

Routine regulatory monitoring for PFAS in consumer goods overwhelmingly depends on target-only analytical methods, while non-target analysis (NTA) remains rarely implemented in official compliance testing.[2,24] Targeted detection can only identify a predefined short list of well-characterized legacy PFAS such as PFOA and PFOS, relying on pre-purchased chemical standards for comparison. However, many emerging replacement PFAS, transformation products, and novel fluorinated analogs fall outside these target lists and therefore escape routine inspection.[1] For instance, total organofluorine measurements of cosmetic samples have revealed substantial fluorine content that cannot be attributed to any target-listed PFAS, demonstrating large pools of unidentified fluorinated ingredients in long-wear makeup marketed to consumers.[25] Similarly, in grease-proof fast-food paper packaging, target-based testing only accounts for less than 6% of total extractable organofluorine, meaning most fluorinated compounds present in packaging remain undetected under conventional laboratory protocols.[13]

The global lack of standardized, widely accessible non-target screening creates two interconnected ethical problems. First, regulatory agencies and consumers receive incomplete data about real-world PFAS exposure, so the public health risks from unlisted substitutes are systematically underestimated.[21] Corporations can deploy new, structurally modified PFAS alternatives that meet existing target-analysis compliance thresholds yet still retain persistence and bioaccumulative potential, creating a regulatory loophole.[1] Second, non-target high-resolution-mass-spectrometry instruments

require expensive hardware, specialized staff, and heavy computational resources; low- and middle-income countries seldom have such capacity, widening the environmental-justice gap between wealthy jurisdictions and resource-limited regions.[2] Even when non-target workflows exist in research laboratories, they are rarely translated into mandatory consumer-product compliance tests, so hidden PFAS contamination in cosmetics, textiles, and food packaging continue to go unreported for ordinary end-users.

3. PFAS in Drinking Water

This chapter discusses the presence of PFAS in drinking water systems, and analyzes the ethical problems aligned with global inconsistency, spatial inequality, and the need for technical progress.

3.1 The Hidden Toxicity in Our "Water Bottles"

Apart from existing in consumer goods, PFAS compounds also migrate into groundwater and surface-water sources, contaminating public and private drinking-water supplies across many regions.[1] This can stem from industrial waste and runoff, such as from a paper production factory,[26] as well as from airports, military bases, and fire training sites where film-forming foams are present. Investigations across the United States have detected measurable PFAS in drinking-water systems serving millions of residents (more than 99% of US residents have long-chain PFAS in their blood),[27] with private household wells in rural communities particularly vulnerable because they often receive no routine regulatory monitoring.[28] Once PFAS enters drinking water, continuous consumption creates chronic low-dose internal exposure; these substances accumulate slowly within human blood and tissues at trace-level concentrations.[29] Pregnant people, infants, and children are especially susceptible groups: prenatal drinking-water-related PFAS exposure has been associated with altered birth outcomes, suppressed immune responses, and developmental harm.[9] Drinking water has become one of the main sources of PFAS exposure (Fig. 2), raising health concerns and also ethical problems that we are going to discuss next.

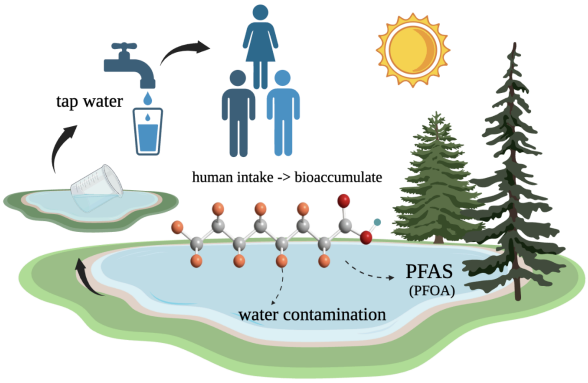


Figure 2: Human intake of PFAS from drinking water

3.2 Global Inconsistency and Spatial Inequality

National standards of PFAS-polluted drinking water show profound global inconsistency, generating severe spatial inequality in public-health protection across different regions.[21] The EU Drinking Water Directive sets a combined total-PFAS limit of 100 ng/L for drinking-water supplies, while the United States Environmental Protection Agency (U.S. EPA) enforces strict maximum contaminant levels of 4 ng/L each for PFOA and PFOS, alongside binding limits for four additional PFAS variants.[19,28] However, over 75% of low- and middle-income regions have no legally binding national drinking-water guideline values for PFAS; many only conduct sporadic target-only monitoring limited to legacy PFOA and PFOS, with no official surveillance of short-chain replacement compounds, driving inequality in human health and environmental conditions.[21] Even within high-income nations, sharp domestic spatial disparities still persist: roughly 13 % of the U.S. population relies on

unregulated private household wells, outside federal drinking-water testing mandates; field surveys have detected measurable PFAS in 63 % of sampled private wells, with up to 49 % exceeding federal health-based benchmarks.[30]

PFAS exposure and protection differ between countries and communities. In parts of Africa, including Ghana and Nigeria, PFAS contamination has been reported, but monitoring and treatment remain limited.[31] PFAS has also been found in water sources in India, but without regular monitoring, some pollution may go undetected.[32] In contrast, the United States and European Union have more established PFAS monitoring systems and stricter drinking-water rules. However, monitoring data also reveal inequalities within countries: a study of 7,873 community water systems across 18 U.S. states found that PFAS detection was positively associated with the proportions of Black and Hispanic residents served.[33,34] This means that people do not receive the same protection from PFAS across countries and communities. This spatial injustice creates an ethical paradox: equivalent toxic risks from PFAS-contaminated drinking water receive vastly different levels of regulatory response purely based on geographic location and socioeconomic status. Populations in resource-limited regions face preventable chronic exposure without official warning or affordable remediation options, even though PFAS pose identical biological hazards for all human populations.[35]

3.3 Lack of Cost-effective Detection and Purification Methods

In addition, conventional detection and purification methods for target PFAS have limitations. For instance, Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS), is the current gold standard for PFAS detection, offers high sensitivity and selectivity but requires high capital investment and well-trained technical staff. Gas Chromatography-Mass Spectrometry (GC-MS) is well-suited for volatile neutral PFAS precursors such as fluorotelomer alcohols, but remains target-oriented and requires skilled operators. Total Organic Fluorine (TOF) analysis allows non-target screening, but it cannot distinguish PFAS from other fluorinated compounds and is fairly costly to use.[36]

Several methods have already been used to remove PFAS from drinking water, including granular activated carbon (GAC), anion exchange (IX), and membrane filtration such as reverse osmosis (RO) and nanofiltration (NF). GAC is widely used but is less effective for short-chain PFAS. IX uses resins to capture PFAS. Different resins vary in their ability to capture PFAS. Some substances in water can also occupy binding sites on the resin, leaving fewer sites available to capture PFAS. RO and NF can remove many PFAS, but they require high pressure and produce a concentrated waste stream. These methods only remove PFAS from water without destroying them. The used carbon, resins, and concentrated wastewater also need further treatment or disposal, adding to the cost.[37] These costs can make it harder for water systems with limited funding to adopt these methods.

In some regions where funding is available, water treatment plants already use these methods to remove PFAS from drinking water: Anaheim, California, operates nine PFAS treatment plants,[38] while Katherine, Australia, operates one permanent plant capable of treating up to 10 million liters of water per day.[39] Building and running these facilities requires substantial funding. California's N Wells IX facility required an initial investment of US$9.6 million and annual operation and maintenance costs of US$1.7 million.[40] For costs per unit of drinking water treated, a Swedish assessment estimated combined capital and operating costs of €0.30 per m^3, including operating costs of €0.18 per m^3.[41] Both the initial investment and

the continuing expense can put PFAS treatment beyond the budgets of communities with limited resources.

Fortunately, researchers are exploring ways to reduce this financial burden. For example, Cheng et al. used flash-encapsulated fluorination to recover approximately 90% of the fluorine from PFAS-loaded activated carbon as reusable silver fluoride.[42] Their progress offers hope that effective PFAS treatment will become more affordable for communities with limited resources.

Therefore, most low- and middle-income nations or regions cannot afford the high cost of the PFAS detection and purification methods, leaving them uncertain about whether the tap water or well water is PFAS contaminated and unable to purify water frequently. This further raises the ethical issues such as global inequality in public health and will also lead to more contamination in the entire environment and ecosystem.

4. Ethical Analysis

This chapter mainly discusses the ethical principles lay behind the issues that PFAS have brought about, including ethical demonstrations of global inconsistency and inequity, content intransparency, and human welfare over generations.

4.1 Distributive Justice

Distributive justice refers to the fair, equitable, and appropriate distribution of benefits and burdens across society, ensuring that no specific demographic is disproportionately subjected to systemic risks or denied fundamental resources.[43]

However, due to the unequal income distribution across some developed countries and developing areas, people in low income regions often drink private well water or tap water without PFAS detection or pretreatment, posing a heavier risk there. Also, many factories or industries that may discharge PFAS-included waste water into the local water system are more willing to stay besides poorer areas, in that they can hire cheaper laborers and don't have some strict legislation or restrictions. This leads to more PFAS contaminated water, especially drinking water, and brings about even more severe potential health threats to people who cannot afford frequent detections or pretreatments. Besides, since there is global legislation inconsistency on legal PFAS added amounts across developed and developing countries, more PFAS products are still being transported to middle- and low- income areas, which furthermore aggravates the injustice towards people.

According to the principle *"Distributive Justice"*, total fairness in how society shares the good and the bad should not depend on their financial abilities or the geographic regions. Good things like clean drinking water and bad things like toxic manufacturing runoff and PFAS-rich consumer goods shouldn't be dumped on one specific group just because they are in a developing country or have less money.

4.2 Respect for Autonomy

Respect for Autonomy is another ethical principle that is defined on the moral obligation to respect the self-determination and decision-making capacities of autonomous persons, ensuring they can act intentionally, with understanding, and without controlling influences.[44] This indicates that people have the right to make their own choices about their lives based on enough transparent information.

For this principle to work, people should be given all the facts for making their own decisions. In this study, specifically, the lack of clear contents of the added amounts of PFAS in consumer goods obviously does not follow this rule. Without being fully informed, many people are just purchasing and using daily goods with a certain amount of PFAS intake unconsciously,

living around potential health risks every day without even knowing their existence.

Therefore, content transparency of various products should be implemented worldwide to ensure respect for autonomy and to protect human rights and health.

4.3 Nonmaleficence

Nonmaleficence is the ethical obligation to not intentionally or needlessly inflict harm on others, rooted in the foundational medical and bioethical maxim *primum non nocere* (translated as *first, do not harm*).[44]

Nonmaleficence requires us to lower or ban any harms that might be otherwise brought to people or the ecosystem in general. If a chemical manufacturer knows a substance causes health issues, this principle states they have a primary duty to stop polluting consumer products or water systems with it. That means with various studies showing the health threats associated with PFAS, industries should not produce any PFAS-contained consumer goods even if they indeed bring a lot of economic benefits and convenience in life.

4.4 The Precautionary Principle

"When an activity raises threats of harm to human health or the environment, precautionary measures should be taken even if some cause and effect relationships are not yet fully established scientifically" (Wingspread Consensus Statement), this is the official definition for *The Precautionary Principle*.[45,46]

This ethical concept particularly stresses the importance of keeping safety in the first place. It demonstrates that if a new substance shows strong signs of being harmful, regulators should restrict it rather than waiting decades for an absolute scientific proof after people have already gotten sick. By applying this to real practice, it means governments should better cut down all the emerging replacement PFAS, transformation products, and novel fluorinated analogs that fall outside the target lists or current regulation rules. It is vital to find the actual risks behind these substances with accurate scientific methods, yet that does not mean we should sacrifice human safety for years before we find an exact answer.

4.5 Intergenerational Justice

Intergenerational justice states the moral obligation to regulate resource consumption, pollution, and systemic risks so as to meet the needs of the present without compromising the ability of future generations to meet their own needs.[47]

In other words, this principle shows that we shouldn't ruin the planet or leave behind toxic chemicals such as PFAS that take centuries to break down, because doing so steals a safe and healthy environment from our children and grandchildren. Instead, in order to better protect our future generations and the whole living planet, we should try our best to lower the potential threats that PFAS poses, which requires the together work of governments, manufacturers, and all human beings.

5. Discussion

Based on the ethical concerns addressed above, this chapter proposes several potential solutions to the problems.

5.1 Class-Wide Restrictions and Source Control Regulatory Models

Regulatory bodies should move beyond compound‑by‑compound target‑based restrictions and adopt class‑wide PFAS bans that also include some hazardous PFAS alternatives or non-essential PFAS organics. Instead of only banning well‑studied legacy substances such as PFOA and PFOS, regulators should apply restrictions to the entire PFAS chemical class for cosmetics, waterproof clothing, and disposable food packaging, except narrow, verified essential‑use cases for medical or firefighting

equipment.[19] This policy approach closes the "regulatory treadmill" loophole in which manufacturers simply replace banned long-chain PFAS for structurally similar short-chain replacements that still carry persistence and toxic potential.[1]

5.2 Community-Led Monitoring and Capacity Building in Testing

NTA should be gradually integrated into official compliance monitoring, rather than remaining confined to academic research laboratories.[2] International aid and collaborative scientific programs should expand affordable NTA laboratory capacity in low- and middle-income countries; without improved testing infrastructure, jurisdictions cannot identify hidden PFAS contamination in imported consumer goods or local drinking-water aquifers.[31] Besides, it is essential to develop point-of-care detection methods that are efficient, easy to operate, and also cost-effective. Cheaper, faster detection and purification methods would allow many more low- and middle-income regions to afford testing, improving drinking water quality and promoting public health equity globally.

5.3 Economic Incentives and International Trade Frameworks

Market-driven transition toward fluorine-free alternatives should also be encouraged. As chemical firms including 3M, Chemours, and Amcor have already proven technically viable fluorine-free coatings for food packaging, cookware, and textiles, policy incentives could accelerate broader adoption of these replacements.[13] Tax incentives and public procurement preferences for fluorine-free consumer goods can offset higher short-term manufacturing costs, avoiding a two-tier market where safe alternatives remain limited to premium-priced high-end products.

To prevent low- and middle-income countries from becoming dumping grounds for non-compliant imported consumer goods, international trade frameworks must integrate border-adjusted compliance standards. Import verification mechanisms should require third-party certification of fluorine-free status for consumer goods arriving at customs, further reducing inequity among PFAS contaminations.

6. Conclusion

In conclusion, PFAS are artificial chemicals that exhibit useful properties for product making as well as resistant degradation that poses long-term health threats and environmental risks. While industrial sources have long been documented, civilian-oriented consumer products and contaminated drinking-water are often overlooked exposure pathways for ordinary people worldwide. This paper has illustrated a wide range of interconnected ethical conflicts including PFAS global legislation inconsistency and spatial inequity, content intransparency, the lack of non-target PFAS analysis, and the need to develop more cost-effective and user-friendly detection and water purification methods.

For future regulations, governments should restrict non-essential uses of PFAS and require factories to reduce PFAS releases. Scientists should develop portable tools for rapid, on-site PFAS detection and make them available in areas with potential PFAS contamination. To protect their farmland, local residents also need to work with research institutions to collect water samples near possible pollution sources. Globally, PFAS pollution is a problem that requires contributions from all countries. Trade in PFAS-containing products can spread pollution across borders. Without effective international rules, companies may move polluting production from high-income countries to developing countries, placing a greater burden on these regions.

Although the Stockholm Convention already controls several groups of PFAS and related compounds, it does not cover the entire PFAS family. Since the PFAS family contains thousands of substances, international controls should cover more PFAS and their precursors. This would help reduce PFAS pollution and the unequal risks faced by different regions.

7. References

[1]Wang, Z., et al. (2017). A never-ending story of per- and polyfluoroalkyl substances (PFASs)? Environmental Science & Technology, 51(5), 2508–2518. https://doi.org/10.1021/acs.est.6b04806

[2]Zhao, L., et al. (2026). Integrating target, nontarget analysis with machine learning to illuminate PFAS characteristics and health risks in Chinese cosmetics. Environmental Pollution, 397, 127933. https://doi.org/10.1016/j.envpol.2026.127933

[3]Souza, M. C. O., et al. (2026). PFAS and One Health: Integrating human, animal, and environmental perspectives. Archives of Toxicology, 100, 1769–1787. https://doi.org/10.1007/s00204-026-04306-1

[4]Lin, X., et al. (2023). Characteristic and health risk of per- and polyfluoroalkyl substances from cosmetics via dermal exposure. Environmental Pollution, 338, 122685. https://doi.org/10.1016/j.envpol.2023.122685

[5]Panieri, E., et al. (2022). PFAS molecules: A major concern for the human health and the environment. Toxics, 10, 44. https://doi.org/10.3390/toxics10020044

[6]Campbell, K. S., et al. (2026). An ecosystem-scale model of PFAS dynamics in stream-to-riparian food webs. Environmental Science & Technology, 60, 11022–11035. https://doi.org/10.1021/acs.est.5c16942

[7]Lewis, R. C., et al. (2015). Serum biomarkers of exposure to perfluoroalkyl substances in relation to serum testosterone and measures of thyroid function among adults and adolescents from NHANES 2011–2012. International Journal of Environmental Research and Public Health, 12(6), 6098–6114. https://doi.org/10.3390/ijerph120606098

[8]Khanashyam, A. C., et al. (2025). Perfluoroalkyl and polyfluoroalkyl substances contamination in dairy: A global perspective. Science of the Total Environment, 1001, 180968. https://doi.org/10.1016/j.scitotenv.2025.180968

[9]Zheng, Q., et al. (2024). The effect of PFAS exposure on glucolipid metabolism in children and adolescents: A meta-analysis. Frontiers in Endocrinology, 15. https://doi.org/10.3389/fendo.2024.1261008

[10]Roth, K., et al. (2022). Exposure to per- and polyfluoroalkyl substances (PFAS) and type 2 diabetes risk. Frontiers in Endocrinology, 13. https://doi.org/10.3389/fendo.2022.965384

[11]Paudel, D., et al. (2025). A scoping review on per- and poly-fluoroalkyl substances (PFAS) and colorectal cancer: Evidence from in vitro, animal, and epidemiological studies. Environment International, 203, 109778. https://doi.org/10.1016/j.envint.2025.109778

[12]Ragnarsdóttir, O., et al. (2022). Dermal uptake: An important pathway of human exposure to perfluoroalkyl substances? Environmental Pollution, 307, 119478. https://doi.org/10.1016/j.envpol.2022.119478

[13]Schaider, L. A., et al. (2017). Fluorinated compounds in U.S. fast food packaging. Environmental Science & Technology Letters, 4(3), 105–111. https://doi.org/10.1021/acs.estlett.6b00435

[14]Namazkar, S., et al. (2024). Characterization and dermal bioaccessibility of residual- and listed PFAS ingredients in cosmetic products. Environmental Science: Processes & Impacts, 26(3), 482–492. https://doi.org/10.1039/D3EM00461A

[15]U.S. Food and Drug Administration. (2025). Report on the use of PFAS in cosmetic products

and associated risks. http://www.fda.gov/media/190319/download

[16]Yang, Y., et al. (2022). Per- and polyfluoroalkyl substances (PFAS) in consumer products: An overview of the occurrence, migration, and exposure assessment. Environmental Science Europe, 34(1), 142. https://doi.org/10.3390/molecules30050994

[17]Xia, C., et al. (2022). Per- and polyfluoroalkyl substances in North American school uniforms. Environmental Science & Technology, 56(19), 13845–13857. https://doi.org/10.1021/acs.est.2c02111

[18]Lendewig, M., et al. (2025). PFAS regulations and economic impact: A review of U.S. pulp & paper and textiles industries. Chemosphere, 377, 144301. https://doi.org/10.1016/j.chemosphere.2025.144301

[19]European Commission. (2024). Commission Regulation (EU) 2024/2462. http://data.europa.eu/eli/reg/2024/2462/oj

[20]TRIS–European Commission. (2025). Notification detail: National draft rules restricting PFAS in consumer articles. https://technical-regulation-information-system.ec.europa.eu/en/notification/26455

[21]Sharma, B. M., et al. (2026). Global disparities in the regulation of PFASs: The risk of shifting the PFAS pollution burden to developing countries. Environmental Science & Technology, 60(11), 8266–8278. https://doi.org/10.1021/acs.est.5c14777

[22]Zou, H., et al. (2023). Continuing large-scale global trade and illegal trade of highly hazardous chemicals. Nature Sustainability, 6, 1394–1405. https://doi.org/10.1038/s41893-023-01158-w

[23]Erickson, B. E. (2021). PFAS undisclosed in most cosmetics sold in the US and Canada. Chemical & Engineering News, 99. https://cen.acs.org/environment/persistent-pollutants/PFAS-undisclosed-cosmetics-sold-US/99/i23

[24]Shao, G., et al. (2025). Target and non-target screening of per- and polyfluoroalkyl substances (PFAS) in drinking water: Focus on alternatives and health risks. Environmental Research, 285, 122581. https://doi.org/10.1016/j.envres.2025.122581

[25]Whitehead, H. D., et al. (2021). Fluorinated compounds in North American cosmetics. Environmental Science & Technology Letters, 8(7), 538–544. https://doi.org/10.1021/acs.estlett.1c00240

[26]Langberg, H. A., et al. (2021). Paper product production identified as the main source of per- and polyfluoroalkyl substances (PFAS) in a Norwegian lake: Source and historic emission tracking. Environmental Pollution, 273, 116259. https://doi.org/10.1016/j.envpol.2020.116259

[27]Heckel, E. J., et al. (2026). Residential history and drinking water type as predictors of serum PFAS in communities with a history of contaminated drinking water. Journal of Exposure Science & Environmental Epidemiology. https://doi.org/10.1038/s41370-026-00937-9

[28]U.S. Environmental Protection Agency. (2026). Per- and polyfluoroalkyl substances (PFAS): Final PFAS national primary drinking water regulation. https://www.epa.gov/sdwa/and-polyfluoroalkyl-substances-pfas

[29]Cousins, I. T., et al. (2016). The precautionary principle and chemicals management: The example of perfluoroalkyl acids in groundwater. Environment International, 94, 331–340. https://doi.org/10.1016/j.envint.2016.04.044

[30]RTI International. (2025). "Forever chemicals" in well water: Study finds private wells commonly exceed PFAS health guidelines. https://www.rti.org/news/private-wells-exceeding-pfas-health-guidelines-research

[31]Adewuyi, A., & Li, Q. (2024). Emergency of per- and polyfluoroalkyl substances in drinking water: Status, regulation, and mitigation strategies in developing countries.

Eco-Environment & Health, 3(4), 355–368. https://doi.org/10.1016/j.eehl.2024.05.008
[32]Mandal, J., et al. (2025). PFAS contamination: Regulatory landscape and future perspectives for India. Proceedings of the Indian National Science Academy. https://doi.org/10.1007/s43538-025-00485-0
[33]Liddie, J. M., et al. (2023). Sociodemographic factors are associated with the abundance of PFAS sources and detection in U.S. community water systems. Environmental Science & Technology, 57, 7902–7912. https://doi.org/10.1021/acs.est.2c07255
[34]Tokranov, A. K., et al. (2024). Predictions of groundwater PFAS occurrence at drinking water supply depths in the United States. Science, 386(6723), 748–755. https://doi.org/10.1126/science.ado6638
[35]Yu, R.-S., et al. (2025). A global overview of per- and polyfluoroalkyl substance regulatory strategies and their environmental impact. Toxics, 13(4), 251. https://doi.org/10.3390/toxics13040251
[36]Chung, Y., et al. (2025). Comparative review of conventional and emerging PFAS detection methods: A focus on detection strategies. Microchemical Journal, 217, 15103. https://doi.org/10.1016/j.microc.2025.115103
[37]Tshangana, C. S., et al. (2025). Technology status to treat PFAS-contaminated water and limiting factors for their effective full-scale application. npj Clean Water, 8(41). https://doi.org/10.1038/s41545-025-00457-3
[38]City of Anaheim Public Utilities. (n.d.). Groundwater treatment program. Retrieved September 28, 2026, from https://www.anaheim.net/5817/Groundwater-Treatment-Progra
[39]Power and Water Corporation. (n.d.). Katherine's water supply. Retrieved September 28, 2026, from https://www.powerwater.com.au/about-us/what-we-do/water-supply/katherine-water-supply
[40]deSilva, V., & Deshmukh, D. (2025). The cost to remove PFAS: A review of US water treatment plants. Journal AWWA, 117(4), 54–63. https://doi.org/10.1002/awwa.2436
[41]EurEau. (2023, September 18). Universal PFAS restriction: Consultation on Annex XV report – Additional EurEau comments. https://www.eureau.org/resources/consultations/7441-eureau-accompanying-document-to-public-consultation-on-upfas/file
[42]Cheng, Y., et al. (2026). Flash-encapsulated fluorination for circular valorization of fluorine waste streams. Nature Chemical Engineering, 3, 535–543. https://doi.org/10.1038/s44286-026-00442-8
[43]Rawls, J. (1971). A theory of justice. Belknap Press. https://doi.org/10.2307/j.ctvjf9z6v
[44]Beauchamp, T. L., & Childress, J. F. (2019). Principles of biomedical ethics (8th ed.). Oxford University Press.
[45]Science and Environmental Health Network. (1998, January 26). Wingspread consensus statement on the precautionary principle.
[46]United Nations. (1992). Rio declaration on environment and development. In Report of the United Nations Conference on Environment and Development (Principle 15).
[47]World Commission on Environment and Development. (1987). Our common future (The Brundtland Report). Oxford University Press.

The Ethics of Incorporating Nanotechnology into Food Packaging and Food Storage Materials

Taylor Maguire[1*], Ariana Lye[2*]
[1]*Garden City High School, NY, 11530,* [2]*Syosset High School, NY, 11791*

Abstract

Food packaging and food storage materials are constantly adapting to better protect food against mechanical forces and environmental factors such as temperature, light, and moisture. Nanomaterials and nanocomposites, innovative materials engineering on the nanoscale (1-100 nm), have garnered attention in recent years for their unique properties as antioxidant, thermostable, and UV-blocking compounds. As such, materials such as metal and metal oxide nanoparticles as well as nanoclays have begun to be used in applications such as food packaging and storage. However, the same material characteristics that afford them their many laudable properties, such as their bioactive and chemical properties, can also pose risks to the same entities they might benefit. This paper seeks to address the underlying ethical concerns surrounding widespread use of nanotechnology in such settings by specifically discussing, in tandem with its benefits, the human health risks, environmental concerns, and government regulations surrounding this newly implemented group of materials. Based on current gaps in available food packaging-related research and regulations regarding the effects of nanomaterial exposure on long-term toxicity, water system contamination, and environmental accumulation, an evaluation is formed. This paper further proposed a recommendation, formulated on ethical grounds, to promptly perform recurring studies on a diverse range of nanomaterial additives, implement limitations in the case of environmental buildup, and require product testing under a broader range of realistic conditions before being available to consumers.

1. Introduction

The food packaging and food storage industries play a key role in feeding Earth's population by protecting food from contamination or spoilage, preventing physical damage during transport, and reducing overall food waste. As of 2025, the worldwide food packaging industry itself was considered to be growing and was estimated to reach a value of 512 billion USD by 2028.[1] This expected growth reflects the fact that food-contacting materials are always evolving to accommodate developments in active and intelligent packaging, which allow for longer shelf life and better protection against environmental stressors.

Recent developments in food packaging and storage include the incorporation of nanotechnology. Coined by Japanese scientist Norio Taniguchi in 1971 after being alluded to by American physicist and Nobel Prize laureate Richard Feynman in 1954, nanotechnology refers to the area of research and engineering that manipulates and controls matter on the nanoscale ranging from 1 to 100 nanometers.[2] While nanomaterials come in a wide variety of shapes and sizes, metal and metal oxide nanoparticles are a widely recognized example. These materials are often combined with existing solids to form nanocomposites, lending their mechanical strengthening, thermal stabilizing, barrier enhancing, antimicrobial, and antioxidant properties to food storage.[3] In this way, nanotechnology has paved the way for the future of material design and could work to improve product shelf life and even reduce global food waste.

However, the increasing popularity and usage of nanomaterials has raised ethical questions regarding their use in food packaging. There are still many dangers associated with nanotechnology, which must be equally considered in order to evaluate the merits of food-contacting nanomaterials. Therefore, this paper seeks to examine the risks associated with nanomaterials, particularly in relation to human health and environmental toxicity, while also acknowledging existing regulations and what can be done to control this growing industry moving forward.

2. Development & Innovations

Nanomaterials have garnered significant attention in the science and technology world ever since the advent of the modern era of nanotechnology in the mid- to late 1900s.[2] Nanomaterials, ranging from carbon-based materials to metal and metal oxide nanoparticles and nanoclays, are largely recognized for their unique properties often attributed to their small size, large surface-area-to-volume ratio, and high reactivity.[4] They can also be incorporated into existing materials to form nanocomposites, new solids made of two or more phases in which one exists at the nanoscale. With wide commercial applications across medicine, energy, electronics, and the environment, nanomaterials are at the forefront of innovation and are frequently adapted for novel purposes.

One controversial industry that experienced the introduction of nanomaterial technology in the early 2000s was the food packaging industry.[5] The primary goal of food packaging, and by extension food storage containers, is to protect food products from agents which have the ability to degrade, break down, or contaminate their contents. Common concerns include, but are not limited to, light, oxygen, moisture, biological organisms, and mechanical forces. To accommodate such factors and increase the quality of stored food, the food packaging industry is constantly evolving. Therefore, engineered nanocomposites, often incorporating metal/metal oxide nanoparticles like zinc oxide (ZnO-NPs), silver (Ag-NPs), and titanium dioxide (TiO_2-NPs) or nanoclays like montmorillonite (MMT) and kaolinite, have made their way into this industry and food storage market in the hope of overcoming existing material limitations.[6]

Innovations in nanomaterials and nanocomposites currently benefit the food packaging industry by improving the physical properties of packaging and storage materials.[7] First and foremost, nanomaterials increase mechanical strength by enhancing both structural networks and solid content, ensuring adequate protection of food during transport, handling, and storage.[8] In the presence of functional groups, nanofillers may experience hydrogen bonding and electrostatic adsorption to form tight polymer matrices. In other instances, additive materials like nanoclays and ZnO nanosheets increase material thickness and, by extension, tensile strength. Secondly, thermal stability has been shown to increase in composites incorporating ZnO-NPs, TiO_2-NPs, nano cellulose, nano chitosan, and nano protein.[9] This property serves to mitigate dangerous chemical and plastic breakdown during microwave heating and the storage of hot substances. Thirdly, nanotechnology improves the physical barrier surrounding foodstuffs, blocking out UV light, moisture, and unwanted gases from the environment.[10] ZnO and TiO_2 nanoparticles, widely recognized for their UV-blocking capability specifically, protect against photodegradation when applied to the food storage barrier. Meanwhile, nanomaterials on the whole reduce gas and moisture diffusion due to their characteristic polymer interactions and induced morphological changes. Reduced oxygen exposure is particularly important, as it can extend the shelf life by slowing oxidation.

Beyond physical enhancements, nanocomposites also have the potential to actively benefit food storage containers due to their chemical and biological properties. Active nanomaterials, unlike passive materials, work by interacting with food and its environment and involve either a direct reaction or change. Antimicrobial and antioxidant activity can be attributed to active nanocomposites due to the presence of functional antimicrobial and oxygen-scavenging agents. These agents fight microbial pathogens and reduce damaging oxidative stress to prevent food spoilage and nutritional breakdown. Metal and metal oxide nanoparticles exhibit antimicrobial activity due to the controlled release of metal ions, with one study specifically illustrating that copper oxide (CuO) reduces viable bacteria, algae, and fungi numbers.[11] Nanoclays, on the other hand, have an antioxidant effect via oxygen scavenging activity. Intelligent packaging and storage is another highly innovative application of nanomaterials in which technology can be applied to. Intelligent packaging materials such as sensors, detectors, and Nano Radio Frequency Identification (RFID) tags look beyond the traditional goal of food storage toward product authenticity and optimization.[12]

Promoting nanomaterial and nanocomposite packaging has implications beyond just food storage itself. By increasing shelf life and spoilage detection, this technology could possibly reduce food waste and postharvest losses as well as increase food quality and longevity.[13] Food waste, the loss of food within the supply chain, is a significant concern globally, as nearly 800 million people go hungry each year and nearly one-fifth of all food available to consumers goes unconsumed, not to mention that it is a significant waste of water, land, and energy, and produces greenhouse gases.[14] Waste mitigation could also lead to greater resource allocation to impoverished communities, feeding millions in need while also saving money. On the other hand, nanomaterial technology could allow for the production of improved biodegradable materials for food packaging.[3] Since nanomaterials are often used to increase the material strength and barrier properties of materials that otherwise couldn't compete with traditional polymers, their integration worldwide may benefit society by decreasing plastic consumption and landfill buildup.

Yet, despite the short-term and long-term benefits of incorporating nanomaterials into food storage containers, their merits within the food packing industry and beyond are not so simple. Ethical considerations regarding the safety of nanomaterials to both human health and Earth's environments, as well as the efficacy of existing regulations, must also be made and addressed in order to further explore nanomaterial applications in food storage.

3. Ethical Considerations

3.1 Human Health Risks

The main ethical concern regarding food packaging is the danger posed by leaching of nanoparticles from the nanocomposites they are contained in to the food, where unsuspecting consumers would then undergo an oral route of uptake into the body.[15] The hazards caused by these nanoparticle-embedded polymers occur mainly in one of two ways: either the particles themselves can physically diffuse into the food to be consumed, or they can react under exposure to certain food conditions and then turn into ions that are easily distributed throughout the food that was in contact with these materials.[16] However, another factor that must be taken into account is the wide variance of toxicity they can have depending on their characteristics. The exact toxicity posed by nanoparticles depends greatly on their specific size and what kind of polymer they are entrapped in, the treatment of the packaging and

what chemical environment it is exposed to when in contact with the food, and what kind of mechanical stress it has been put under.[17] Most notably, according to calculations where nanoparticles are modeled as molecules, the diffusion that the particles undergo into the food rapidly decays from the range of particles sized 1 nm to 4 nm.[17] The size range of nanoparticles is 1-100 nm, and by 4 nm, there is practically no diffusion. In the most hazardous possible host polymer with the smallest particle size, this calculation indicates that the concentration of nanoparticles that diffuse into the food can exceed levels that have been shown to cause decline in cellular function in the food itself, but at concentrations to which they would end up diluted in the body, leaching would incur a negligible negative cellular effect.[18]

In terms of the mechanisms of cellular toxicity at actually toxic levels, large trends in mechanisms for toxicity are also fairly consistent across different nanoparticles and vary more with size even in the face of differences in chemical identity. The most typical method of cellular impairment is the creation of reactive oxygen species.[19] This can occur either when metal or metal oxide nanoparticles are gradually dissolved after arriving at the cell, forming metal ions or prior if they react when in contact with the food, forming ions that make their way into the body leading to the acceleration of the ROS forming Fenton reaction; this leads into the common mechanism of ROS being tied to mitochondria linked accelerated programmed cell death.[20] Additionally, further tied to their size is the ability for some nanoparticles to cross the nuclear membrane, leading to interaction with and damage of the DNA.[21]

Overall, the two largest concerns for toxicity are the reactions with nanoparticles caused by exposure to extreme environments, such as low pH causing the formation of metal ions, in addition to exposure to the nanoparticles in their physical form caused by abrasions. Hence, while current experimental data and simulations determine that the levels of everyday exposure from simple leaching under static conditions are low enough to not have immediately large effects on consumers,[18] the potential for negative effects is not entirely known due to a lack in varied testing conditions. This lack includes abrasion testing, exposure over time to basic and acidic environments and chronic studies where the effects of low doses exposed over long periods of time are determined.[22,23]

3.2 Environmental Concerns

While nanomaterials and nanocomposites have been involved in the fabrication of biodegradable materials, it is imperative to acknowledge the environmental impacts and contamination risks of nanomaterials themselves, in addition to how nanoparticle leaching relates back to human populations, before promoting their usage. In particular, it is important to consider metal and metal oxide nanoparticles such as silver (Ag-NPs), titanium dioxide (TiO_2-NPs), and zinc oxide (ZnO-NPs) since they are both frequently used in the food industry and commonly found in the environment.

The ecological consequences of nanomaterial contamination stem from their leaching into the surrounding environment during the material's lifespan. One of the most significant causes of environmental release occurs during the disposal of food packaging as trash.[15] When food packaging and storage containers accumulate in landfill deposits, harmful nanoparticles incorporated within the material are released into the soil and/or groundwater. However, the leaching does not stop there: nanoparticles can also be released during material production in the form of raw materials or nano-enabled products and during material use itself.[24] As such, measures must be

taken to regulate both the synthesis process and proper storage and disposal of nanomaterial food packaging. Methods should also be tailored to specific nanomaterials and their release mechanisms, as, for example, nanoclays and metal nanoparticles do not have the same exact properties and stability.

A prominent ethical concern regarding nanomaterial release is bioaccumulation. As nanoparticles persist in the environment, they can make their way into the food chain and food webs due to their incredibly small size.[25] This property allows them to be absorbed by or attach to primary producers, like algae or smaller organisms, before being passed to predators by consumption. Over time, accumulation reaches higher trophic levels and can biomagnify in top consumers like humans. Specific effects of nanoparticles on humans include cell death, genotoxicity, oxidative stress, and inflammatory reactions.[26] In various instances, TiO_2-NPs and Ag-NPs have also been found to accumulate in the liver, intestine, gills, and muscles of fish or within invertebrates exposed to the nanomaterial.[27] What makes bioaccumulation ethically important is that it is often undetectable, with nanomaterials building up inside organisms faster than they can be broken down or disposed of. They can impact both local products and prey as well as commercially producing industries. This can lead to unexpected risks to innocent people and communities while also causing involuntary harm to animals.

Water systems may also be severely impacted by nanomaterial usage in food packaging. Improper waste disposal and flow from landfills result in water pollution. For instance, landfill leachate, a heavily contaminated liquid or material, is often produced as a result of water infiltration followed by chemical, microbial, and physical breakdown and overflows to contaminate its surroundings.[28] Such pollution can have an extreme damaging effect on aquatic ecosystems and the organisms that reside within them. In particular, small nanomaterials and nanoparticles like Ag-NPs and ZnO-NPs have been found to cause oxidative stress, inflammation, gene toxicity, and damage to cell membranes in aquatic organisms, like fish, as they adhere to their skin and gills before migrating into biological tissue.[29] Metal nanoparticles have also been indicated to disrupt aquatic behaviors such as feeding, swimming, and predation. Most importantly, small aquatic organisms tend to have lower survival and reproduction rates and even reproductive failure in the face of high exposure rates. Due to their important role within their individual ecosystems, this can promote biodiversity loss and destabilization of the food chains within their same habitat.[30]

Yet, water pollution can be considered a double-edged sword as it has been suggested to disrupt water purification, leading to exposure in human drinking water. Nanomaterial contamination makes the water purification process more difficult since nanoscale size allows for particles to slip past traditional filtration and sedimentation treatment barriers, thereafter remaining hard to detect and capture.[31] When systems are not equipped to handle such pollutants, nanomaterials can reach drinking water and eventually the consumer. Purified water is already a highly limited resource, with only about 3% of Earth classified as freshwater and 0.5% available for use.[32] The ethical dilemma remains: can this planet afford another widespread source of pollution and potential harm?

Plants are also in danger of nanoparticle uptake via soil and air pollution resulting from food packaging. With regard to soil, nanomaterials enter through the root system, intruding through pores or water channels.[33] Air exposure, by contrast, occurs through openings in plant leaves known as stomata or through cracks in the plant itself. These methods of

contamination yield adverse effects in the form of phytotoxicity, bioaccumulation as previously mentioned, and continued pollution.[34] Phytotoxicity is an especially pressing consideration, as it manifests in the form of cell membrane damage, disrupted nutrient balance, reduced growth/germination, and oxidative stress. In the long term, it can reduce crop yields and the global food supply as well as upend ecosystems that depend on such producers for survival. Moreover, metal nanoparticles directly affect terrestrial environments beyond plants themselves by reducing soil nutrient content, microbial distribution, and fertility as a function of bacterial and fungal death.

Overall, no area of the planet is safe from nanomaterial contamination. Yet, this does not mean that nanomaterials should be avoided altogether. Their nanoscale size and large surface area to volume ratio serve as both a blessing and a curse. The very thing that gives nanoclays, composites, particles, and so on their unique capability to enhance materials and which makes them attractive in the food packaging industry also carries risks to the environment. Size is one widely recognized factor that complicates nanomaterial usage, with smaller size associated with greater nanotoxicity.[15] However, this size dependency also means that controls could be made, in accordance with further research, to maintain a comparatively larger size and reduce negative environmental effects. Similarly, dose and structure determine the toxicity and impact of nanomaterials, providing a gateway for future limitations in the hopes of ensuring environmental conservation. Innovation is always continuing; therefore, advancements should continue to be made to adapt to emerging concerns about nanotechnology in order to find an ethical middle ground that does not sacrifice safety for short-term goals.

3.3 Current regulations

Due to the rapid advancement of the field, current regulations struggle to classify and keep up with the inclusion of nanoparticles in food contact surfaces. This leads to potential ethical implications regarding gaps in safety as new materials are synthesized and companies attempt to rapidly push them into the market.

In the EU, regulations surrounding nanoparticles are restrictive around their size, defining a nanomaterial as something where 50% of that material is made up of particles between 1-100 nm; however, some articles[35] point out that this leads to a lack of regulation around particles with very similar properties that fall just outside either side of this range. This unnuanced definition highlights the lack of regulation as it is rooted primarily in physical attributes as opposed to the biological and chemical effects a material could be lent by nanoscale features. Conversely, there are otherwise strict regulations of materials that fall into this category, only allowing very specific materials to be used in nanocomposites in contact with food as defined in Annex I (Regulation (EU) No 10/2011). By having a strictly defined list of materials that is very difficult to add to, this slightly mitigates risks to consumers but also risks stagnating the promising field. Regarding the introduction of novel nanomaterials, they follow guidelines to determine toxicity that, notably in the case of nanomaterials, take care to limit the maximum dosage to prevent unrealistic property-changing agglomeration to less than 100 μg/mL and additionally prevent overdosing to prevent experimental bias by limiting the dosage to under 50 mg/kg.[36]

In the United States, the governing body over the matter of what materials can be included in food contact surfaces is the FDA. Similar to the EU, a nanomaterial is also legally defined as a material that is between the dimensions of 1-100 nm, but covering a slightly broader range of materials. The US also forces

products that rely on engineered nanoscale features to stand up to the same standards they hold more rigidly defined nanoscale materials to. By doing this, a larger portion of materials that exhibit the same kinds of toxicities caused by interactions typical of nanoscale are held up to the same rigorous testing. This comes from one of the most notable pieces of legislation regarding nanomaterials in the US. It was initially not directly regarding the use of nanomaterials in food contact surfaces, but rather their use in cosmetics. It was established following the lawsuit in Dec. 2011, International Center for Technology Assessment (ICTA), et al. v. Margret A. Hamburg M.D.[37] The need for this lawsuit, however, is an indication of a resistance to altering regulation to accommodate the dangers of upcoming technology. After failing to address a 2006 petition to enact stricter regulations surrounding nanomaterials, a lawsuit was filed and subsequently won, leading to the FDA, unlike the EU, holding materials that did not necessarily fall within the size range but did rely on nanoscale dimensions to exhibit their engineered properties to the same testing as other nanomaterials; it also disrupted regulations that protected macroscale products such as the Generally Recognized as Safe (GRAS) regulation, where a material's macroscale components being safe implicating a product's safety no longer applied to nanoscale materials.[38] Currently, the legislation that denies nanomaterials GRAS regulations also applies to nanomaterials used in food packaging. Furthermore, in the US, the many nanomaterials that are used as antibacterials are also classified as pesticides and are further subject to evaluation by the EPA under the Federal Insecticide, Fungicide, and Rodenticide Act.[39]

Overall, there are many legal similarities in these two regulating bodies with addressing nanomaterials being used in food contact surfaces. They share similar working definitions of what qualify as nanomaterials and must be regulated as such, such as defining them with the same dimensions. However, their main point of contention is also in the context of this categorization. They differ slightly with how they deal with substances that may share some of the dangerous effects of nanoparticles without the exact dimensional qualifications. While the EU attempts to control exceptions with the catch all of only having a small list of allowed substances and then carefully reviewing other substances that can join that list, the US mitigates this risk by including these exceptions within the category of nanomaterials with their additional inclusion of particles engineered to have similar effects.

3.4 Ethical Evaluation

Ultimately, it can be said that, despite the aforementioned ethical concerns, nanomaterials should not be excluded from the food packaging industry. In fact, nanomaterials, composites, and particles represent an emerging frontier in scientific innovation with the potential, in this context, to revolutionize food shelf life, industry sustainability, and intelligent storage. Current release and exposure levels remain low when incorporated into existing polymer matrices, with most researchers indicating that nanomaterials do not at the moment pose an immediate danger as acute impacts are generally correlated with high concentrations.[40]

However, the ethical dangers of nanotechnology in regard to health and the environment are simultaneously valid. Nanoparticles and nanomaterials are infamous for their adverse health effects, ranging from inflammation and oxidative stress to genotoxicity and neurotoxicity. The US Occupational Health and Safety Administration (OSHA) recognizes that nanomaterials can pose distinct health and safety hazards depending on physical and chemical characteristics such as particle size, distribution, agglomeration, shape,

chemical composition, and surface chemistry.[41] Moreover, nano-scale substances have been linked to environmental dangers including bioaccumulation, phytotoxicity, aquatic toxicity, and fresh water contamination. These concerns should not be entirely ignored just because the material was incorporated into nanocomposites when found in FCMs as opposed to being simply readily available.

Current studies admit that there is still a significant knowledge gap surrounding nanomaterial release and exposure levels in the context of food packaging. To complicate matters further, the nano-scale packaging industry itself is expanding; one study predicts that the global nano-enabled packaging market, now valued at US $27.7 billion, may reach US $69.8 billion by 2033.[42] This growing popularity could raise current release levels beyond expected rates, contributing to ethical calculations.

Another aspect that must also be taken into consideration are the ethical implications of unknowns, most notably the unknown of the effects of long term exposure. Because the inclusion of nanomaterials in food sciences is a recent and promising innovation, there has not been enough time nor urgency with regards to studying the chronic effects of nano-particle exposure at these levels through an oral route.[23] Hence it is simply not yet understood whether harm is being caused to people with the continued use of these products. The answer to this question would have huge ethical implications for the product as, if it was known to cause harm and was still being sold, this would be a huge violation of ethics. Additionally, the significance to the answer to this question is only compounded by the aforementioned growing industry and prevalence in consumer goods.

With respect to other unknowns, there is also little understanding of the effects of nanomaterials from an environmental standpoint. It has been long understood that exposure to massive amounts of nanoparticles, for instance those disposed of after industrial scale production of goods containing nanocomposites, causes great harm to surrounding ecosystems, particularly aquatic ones.[43] Contamination of such magnitudes has been, for the most part, rectified, but there remains the question and issue of the effects of environmental leaching occurring on a smaller scale. This must be considered because FCMs are frequently disposed of and end up sitting in landfills, then from there undergo harsh conditions such as a wide range of temperatures UV exposure known to liberate nanoparticles from nanocomposites as the polymer breaks down.[44] Similarly to the case of the other unknown, if the disposal of nanocomposite FCMs harms ecosystems in similar ways to large scale dumping of nanoparticles, then this would also present a large ethical concern due to the vast prevalence of the effects. So this, too, is a question that, for the sake of ethics, must be answered.

Current regulations serve an important role in maintaining the ethics of this industry, with it putting checks on these products that protect the environment and consumer wellbeing. They are currently sufficiently thorough, given what is known of the dangers today, with what deserves scrutiny as nanomaterials. Nonetheless, some alterations can be made to further bolster the ethical merit of these regulations. Particularly, as more is discovered in the ways in which nanoparticles can be released from their matrices, such as under what conditions or treatments and under what timeframes, regulations must remain fluid and adapt to these newly determined dangers.

4. Conclusions

As of now, the inclusion of nanomaterials into the food storage industry sits in an ethical middle ground with compelling

arguments both for and against continued integration. The implementation of nanocomposite materials can serve as a cheap, affordable, and reliable method of keeping food safe while also improving the quality of the food itself. Furthermore, with respect to human health, nanomaterials can actually be beneficial in that they act as antimicrobial agents, preventing the growth of harmful bacteria. However, exposure also incurs negative effects on health at high levels and small particle sizes due to various manifestations of nanotoxicity. In terms of environmental effects, nanotechnology holds the potential to improve packaging by lengthening shelf life and reducing food waste, thereby decreasing the emissions, energy, and land spent on a currently under-consumed resources; conversely, their inclusion in single-use materials that are frequently disposed of may lead to leaching into the environment, a phenomenon which serves as a detriment to aquatic ecosystems, water systems, and plant health.

The key to these problems is proper regulations and laws to maintain an equilibrium with a net good. Unfortunately, there is also evidence that current regulations are slow to adapt to new discoveries in potential risks posed by rapidly advancing technology. Because of this lack of legislative foresight, pushes should continue to be made within the scientific community to better understand the long term negative effects so that this promising technology can be utilized to its fullest potential and legal regulations may be enacted based on a more complete understanding of the limits of their use. One recommended action is that, because the main danger to human health lies in contamination of food from the particles, testing including more realistic conditions, notably ones that may have impacts on the particle exposure, should be mandated. These include testing under abrasive conditions and pH extremes. Furthermore, it should be emphasized that there is a dire ethical need for testing of effects of small doses over great lengths of time as exposure to these products as this is currently unknown and, given the current growth in popularity, the population's exposure will only increase in frequency with time.

5. References

[1] Statista Research Department. "Food packaging industry worldwide - statistics & facts." *Statista*, 17 dec 2025, www.statista.com/topics/10806/food-packaging-industry/?srsltid=AfmBOoqaaRM56vCIt4_zAwyzkZF9ypf85a8dEIt6IHjrl9Y2l6Saz7KC. Accessed 27 Aug. 2026.

[2] Bayda, Samer et al. "The History of Nanoscience and Nanotechnology: From Chemical-Physical Applications to Nanomedicine." *Molecules (Basel, Switzerland)* vol. 25,1 112. 27 Dec. 2019, doi:10.3390/molecules25010112

[3] Ashfaq, Alweera, et al. "Application of Nanotechnology in Food Packaging: Pros and Cons." *Journal of Agriculture and Food Research*, vol. 7, Mar. 2022, p. 100270, https://doi.org/10.1016/j.jafr.2022.100270.

[4] "Nanomaterials." *National Institute of Environmental Health Sciences*, www.niehs.nih.gov/health/topics/agents/sya-nano. Accessed 27 Aug. 2026.

[5] Tsagkaris, Aristeidis S et al. "Nanomaterials in food packaging: state of the art and analysis." *Journal of food science and technology* vol. 55,8 (2018): 2862-2870. doi:10.1007/s13197-018-3266-z

[6] Chaudhary, Pallavi et al. "Relevance of Nanomaterials in Food Packaging and its Advanced Future Prospects." *Journal of inorganic and organometallic polymers and materials* vol. 30,12 (2020): 5180-5192. doi:10.1007/s10904-020-01674-8

[7] Garcia, Coralia V. "Editorial: Nanomaterials in food packaging." *Frontiers in nutrition* vol. 9

1083185. 15 Nov. 2022, doi:10.3389/fnut.2022.1083185
[8] Peng, Bo, et al. "Recent Advances in Nanomaterials-enabled Active Food Packaging: Nanomaterials Synthesis, Applications and Future Prospects." *Food Control*, vol. 163, Sept. 2024, p. 110542, https://doi.org/10.1016/j.foodcont.2024.110542.
[9] Zhao, Zhao, et al. "Types of Nanomaterials Commonly Used in Food Packaging, Film Formation Techniques, and Recent Advances in Their Applications." *International Journal of Food Science and Technology*, vol. 60, no. 1, Jan. 2025, https://doi.org/10.1093/ijfood/vvae036.
[10] Mafe, Alice Njolke, et al. "Investigating Nanomaterials for Improved Barrier Properties in Food Packaging." *Al-Mustaqbal Journal of Sustainability in Engineering Sciences*, vol. 2, no. 2, 13 Jan. 2025, https://doi.org/10.62723/2959-5932.1018.
[11] Muthu, Arjun et al. "Nanomaterials for Smart and Sustainable Food Packaging: Nano-Sensing Mechanisms, and Regulatory Perspectives." *Foods (Basel, Switzerland)* vol. 14,15 2657. 29 Jul. 2025, doi:10.3390/foods14152657
[12] Mlalila, Nichrous, et al. "Transformation of Food Packaging from Passive to Innovative via Nanotechnology: Concepts and Critiques." *Journal of Food Science and Technology*, vol. 53, no. 9, Sept. 2016, pp. 3395-407, https://doi.org/10.1007/s13197-016-2325-6.
[13] Obayomi, Oluwatobi Victoria, et al. "Recent Advance in Nanotechnology for Circular Food Systems: Innovations for Sustainable Production, Waste Valorization, and Nutrient Recovery." *Journal of Hazardous Materials Advances*, vol. 21, Feb. 2026, p. 101023, https://doi.org/10.1016/j.hazadv.2026.101023.
[14] "World squanders over 1 billion meals a day - UN report." *UN Environment Programme*, 27 Mar. 2024, www.unep.org/news-and-stories/press-release/world-squanders-over-1-billion-meals-day-un-report. Accessed 27 Aug. 2026.
[15] Gupta, Rakesh Kumar, et al. "Investigating the Toxicological Effects of Nanomaterials in Food Packaging Associated with Human Health and the Environment." *Journal of Hazardous Materials Letters*, vol. 5, Nov. 2024, p. 100125, https://doi.org/10.1016/j.hazl.2024.100125.
[16] R. J. B. Peters and E. J. Hoekstra, in Food Contact Materials Analysis: Mass Spectrometry Techniques, ed. M. Suman, The Royal Society of Chemistry, 2019, pp. 226-244.
[17] Stuparu-Cretu, Mariana, et al. "Metal Oxide Nanoparticles in Food Packaging and Their Influence on Human Health." *Foods*, vol. 12, no. 9, 2023, article 1882, https://doi.org/10.3390/foods12091882.
[18] Awashra, Mohammad, and Piotr Młynarz. "The Toxicity of Nanoparticles and Their Interaction with Cells: An In Vitro Metabolomic Perspective." *Nanoscale Advances*, vol. 5, no. 10, 2023, pp. 2674–2723, https://doi.org/10.1039/d2na00534d. https://doi.org/10.1016/j.scitotenv.2025.181340
[19] Fu, Peter P., et al. "Mechanisms of Nanotoxicity: Generation of Reactive Oxygen Species." *Journal of Food and Drug Analysis*, vol. 22, no. 1, 2014, pp. 64–75. https://doi.org/10.1016/j.jfda.2014.01.005.
[20] Marchi, Saverio, et al. "Mitochondria-ROS Crosstalk in the Control of Cell Death and Aging." *Journal of Signal Transduction*, vol. 2012, 2012, article 329635, https://doi.org/10.1155/2012/329635.
[21] Encinas-Gimenez, Miguel, Pilar Martin-Duque, and Ana Martín-Pardillos. "Cellular Alterations Due to Direct and Indirect Interaction of Nanomaterials with Nucleic Acids." *International Journal of Molecular Sciences*, vol. 25, no. 4, 2024, article 1983, https://doi.org/10.3390/ijms25041983.
[22] Priyadarshi, R., Khan, A., Riahi, Z., & Rhim, J.-W. (2025). Toxicity and safety of nanomaterials in food packaging applications: A

dimension-based review. Comprehensive Reviews in Food Science and Food Safety, 25(1), Article e70374. https://doi.org/10.1111/1541-4337.70374
[23] Meng, X., & Nag, R. (2026). Human health risk assessment of nanoparticles through food consumption — occurrence, exposure, and toxicological implications. Science of The Total Environment, 1013, Article 181340.
[24] Bundschuh, Mirco et al. "Nanoparticles in the environment: where do we come from, where do we go to?." *Environmental sciences Europe* vol. 30,1 (2018): 6. doi:10.1186/s12302-018-0132-6
[25] Dinda, Subas Chandra. 'Bioaccumulation and Toxic Profiling of Nanostructured Particles and Materials'. Unraveling the Safety Profile of Nanoscale Particles and Materials - From Biomedical to Environmental Applications, InTech, 21 Mar. 2018. Crossref, https://doi.org/10.5772/intechopen.74802.
[26] Xuan, Lihui, et al. "Nanoparticles‐Induced Potential Toxicity on Human Health: Applications, Toxicity Mechanisms, and Evaluation Models." *MedComm*, vol. 4, no. 4, 2023, article e327, https://doi.org/10.1002/mco2.327.
[27] Suárez-Oubiña, Cristian, et al. "Bioaccumulation and Human Risk Assessment of Inorganic Nanoparticles in Aquaculture Species." *Environmental Science: Nano*, vol. 11, no. 7, 2024, pp. 2937-47, https://doi.org/10.1039/d4en00167b.
[28] Abdel-Shafy, Hussein I., et al. "Landfill Leachate: Sources, Nature, Organic Composition, and Treatment: An Environmental Overview." *Ain Shams Engineering Journal*, vol. 15, no. 1, Jan. 2024, p. 102293, https://doi.org/10.1016/j.asej.2023.102293.
[29] Subaramaniyam, Udayadharshini et al. "Effects of microplastics, pesticides and nano-materials on fish health, oxidative stress and antioxidant defense mechanism." *Frontiers in physiology* vol. 14 1217666. 26 Jun. 2023, doi:10.3389/fphys.2023.1217666
[30] Mustafa, Sanaa Abdulaziz, et al. "Effect of Environmental Pollutants on Fish Health: An Overview." *Egyptian Journal of Aquatic Research*, vol. 50, no. 2, June 2024, pp. 225-33, https://doi.org/10.1016/j.ejar.2024.02.006.
[31] Kholopo, Mota, and Phoka C. Rathebe. "Lifecycle, Toxicology, and Exposure Pathways of Functional Nanomaterials in Water Treatment: Implications for Environmental and Human Health." *Sustainable Environment*, vol. 11, no. 1, 17 Sept. 2025, https://doi.org/10.1080/27658511.2025.2558248.
[32] Central California Area Office. "Water Facts - Worldwide Water Supply." *United States Bureau of Reclamation*, 4 Nov. 2020, www.usbr.gov/mp/arwec/water-facts-ww-water-sup.html. Accessed 27 Aug. 2026.
[33] Xu, Xinglong, et al. "Nanoparticle-plant Interactions: Uptake, Transport, Physiological Effects, and Environmental Implications." *Journal of Hazardous Materials Advances*, vol. 22, May 2026, p. 101123, https://doi.org/10.1016/j.hazadv.2026.101123.
[34] Islam, Shahidul. "Toxicity and Transport of Nanoparticles in Agriculture: Effects of Size, Coating, and Aging." *Frontiers in Nanotechnology*, vol. 7, 24 June 2025, https://doi.org/10.3389/fnano.2025.1622228.
[35] Franz, Roland, Johannes Bott, and Angela Störmer. "Considerations for and Guidance to Testing and Evaluating Migration/Release of Nanoparticles from Polymer Based Nanocomposites." *Nanomaterials*, vol. 10, no. 6, 2020, article 1113, https://doi.org/10.3390/nano10061113.
[36] EFSA Scientific Committee (Simon More, et al.). "Guidance on Risk Assessment of Nanomaterials to Be Applied in the Food and Feed Chain: Human and Animal Health." *EFSA Journal*, vol. 19, no. 8, 2021, article e06768, https://doi.org/10.2903/j.efsa.2021.6768.
[37] Bergeson, Lynn L. "Sustainable Nanomaterials: Emerging Governance Systems." *ACS Sustainable Chemistry & Engineering*, vol.

1, no. 7, 2013, pp. 724–730, https://doi.org/10.1021/sc4000863.
[38] United States, Food and Drug Administration. *Guidance for Industry: Safety of Nanomaterials in Cosmetic Products*. U.S. Department of Health and Human Services, June 2014, https://www.fda.gov/regulatory-information/search-fda-guidance-documents/guidance-industry-safety-nanomaterials-cosmetic-products.
[39] United States, Environmental Protection Agency. "Pesticide Registration Manual: Chapter 4 - Additional Considerations for Antimicrobial Products." *US EPA*, 1 Mar. 2024, https://www.epa.gov/pesticide-registration/pesticide-registration-manual-chapter-4-additional-considerations.
[40] Gupta, Rakesh Kumar, et al. "Nanotechnology: Current Applications and Future Scope in Food Packaging Systems." *Measurement: Food*, vol. 13, Mar. 2024, p. 100131, https://doi.org/10.1016/j.meafoo.2023.100131.
[41] "Nanotechnology." *Occupational Safety and Health Administration*, U.S. Department of Labor, www.osha.gov/nanotechnology/health-effects. Accessed 10 Sept. 2026.
[42] Patil, Amol. "Nano-enabled Packaging Market Size, Share, and Growth Forecast, 2026 - 2033." *Persistence Market Research*, May 2026, www.persistencemarketresearch.com/market-research/nano-enabled-packaging-market.asp. Accessed 10 Sept. 2026.
[43] Bystrzejewska-Piotrowska, G., Golimowski, J., & Urban, P. L. (2009). Nanoparticles: Their potential toxicity, waste and environmental management. Waste Management, 29(9), 2587–2595. https://doi.org/10.1016/j.wasman.2009.04.001
[44] Reijnders, L. "The Release of TiO2 and SiO2 Nanoparticles from Nanocomposites." Polymer Degradation and Stability, vol. 94, no. 5, May 2009, pp. 873-76, https://doi.org/10.1016/j.polymdegradstab.2009.02.005.

Journal of Ethics in Scientific and Technological Innovations

Ethical and Sustainability Challenges in Modern Agricultural Production

Justin (Jiahong) Ren[1*]
[1]*Shanghai Starriver Bilingual School, Shanghai, China*

Abstract

Modern agriculture has played an essential role in supporting the human population, but its development has also created significant environmental, ethical, and sustainability challenges. This paper examines these challenges from two perspectives: crop production and animal husbandry. The crop-production section discusses the environmental impacts of continuous cropping and the use of fertilizers and pesticides, as well as potential solutions including intercropping, controlled-release fertilizers, exopolysaccharides, and biological control. The section examines ethical concerns related to livestock farming, animal welfare, and artificial selection, and analyzes the challenges to achieving a more sustainable and ethical future for animal husbandry. A more sustainable and ethical agriculture remains limited by technical, economic, and social barriers.

1. Introduction

Agriculture has been a classic example of humans exploiting their surroundings for their own benefit. Throughout history, the development of agriculture was closely associated with major human demographic expansion. A global genetic analysis found that agriculture facilitated a fivefold increase in population growth compared with earlier hunter-gatherer expansions.[1] Agriculture contributes to climate change and environmental degradation through several mechanisms. It occupies roughly 38% of the Earth's land surface, which fuels deforestation. It also accounts for 30–35% of greenhouse gas emissions, such as methane and nitrous oxide. Additionally, agriculture exacerbates biodiversity loss, as well as soil and water degradation. Those issues are expected to worsen with time because, with roughly one in seven people lacking access to food or being chronically malnourished today, the current level of food production is still unable to feed the global population, which is expected to keep growing over the next 50 to 60 years.[2,3] Therefore, it is important to develop and adapt more sustainable practices of agricultural production. This paper examines major sustainability and ethical challenges associated with modern agricultural production, including those associated with continuous cropping, external chemical inputs, livestock farming, animal welfare and artificial selection. This paper also examines alternative practices that can potentially mitigate these challenges.

2. Sustainability Challenges Associated with Crop Production Practices

2.1 Continuous cropping

Agriculture involves the cultivation and production of various crops, and various agricultural practices have been developed and adopted to increase crop yield to feed the global population. Meanwhile, these practices raise ethical concerns because their benefits to humans may come at the expense of other organisms and ecosystems, on which both humans and other species depend. For example, continuous cropping refers to the practice of planting the same crop or its related species continuously on the same plot. Generally, continuous cropping can alter the soil environment by reducing nutrients and soil enzyme activity, causing the imbalance of nutrient element proportions, disturbing microbial populations, increasing pests and

diseases, etc.[4,5] These changes make continuous cropping unsustainable in the long term. Additionally, they can inhibit future crop growth, potentially prompting further human intervention (e.g., increased fertilizer applications or increased cropping area) that may lead to further environmental problems. Continuous cropping is therefore a classic example of soil being treated merely as a means for crop production rather than a complex natural reservoir that supplies nutrients, sustains biodiversity, and performs important ecological functions. Ultimately, the ethical question raised by continuous cropping is whether humans are justified in causing harm to other species and ecosystems in pursuit of agricultural productivity of particular crops. The following case studies will exemplify the mechanisms by which continuous cropping disturbs the environment and how such disturbance can potentially be alleviated.

2.1.1 Case study: continuously cropped tobacco

Tobacco is an economically important crop that has long been continuously cropped worldwide, and it is widely studied as a significant model plant for research purposes.[6] A study of soils subjected to different durations of continuous tobacco cropping found that prolonged continuous cropping was associated with a decline in available phosphorus and an increase in the potassium content in soil, which could contribute to the imbalance of nutrient element proportions. The environmental impacts of tobacco monoculture also expand beyond direct chemical interactions with soil to more complex and significant ones involving soil microorganisms. Research has shown that prolonged continuous cropping of tobacco was associated with the accumulation of phenolic acids and a significant decrease in soil pH. Acidified soil provides suitable breeding conditions for pathogens, deteriorating the soil environment and inhibiting the growth of tobacco and other crops as well.[5]

2.1.2 Case study: continuously cropped garlic

Garlic exemplifies another possible effect of continuous cropping on the environment—altered soil enzyme activity. Soil enzymes are extracellular enzymes produced by soil microorganisms that contribute to soil biochemical processes. A study comparing garlic soils affected by continuous-cropping obstacles with healthy control soils found that the activities of various enzymes in soil (e.g., urease, catalase, sucrase, and polyphenol oxidase) are lower in soil where garlic is continuously cropped. The decreased soil enzyme activity might be related to the acidification of soil associated with garlic monoculture. Additionally, since soil enzymes are mainly produced by soil microorganisms, their decreased activity might also be the consequence of microbial communities being disrupted by continuous cropping, as exemplified in the tobacco case study.[7] Decreased soil enzyme activity can impair important soil biochemical processes that decompose organic compounds into nutrients, reducing soil fertility.

2.1.3 Case study: Garlic/Cucumber Relay Intercropping

The tobacco and garlic case studies demonstrate how prioritizing the continuous production of a commercially valuable crop can alter the soil environment in ways that affect organisms and ecological processes beyond the cultivated crop itself. However, those consequences can potentially be alleviated without completely abandoning continuous cropping. By increasing crop diversity, intercropping provides a solution to the problems associated with continuous cropping. One example is garlic/cucumber relay intercropping, in which garlic and cucumber are

cultivated in the same field during partially overlapping growing periods. A study investigating garlic/cucumber relay intercropping found that compared to cucumber continuous cropping, garlic/cucumber relay intercropping increased the bacterial diversity indices, enhanced soil enzyme activities, improved the availability of nutrients including N, P, and K, and increased the conversion of soil from a fungal to a bacterial type by increasing the number of soil bacteria and decreasing the number of soil fungi, which can potentially replenish the fertility of soil. Eventually, relay intercropping not only alleviates the damage of continuous cropping on the soil environment but also, in turn, promotes cucumber plant growth compared to long-term monoculture,[8,9] which can potentially reduce the need for external human interventions such as fertilization.

2.1.4 Case Study: Maize/Soybean Relay Intercropping

Another example of relay intercropping that produces reduced environmental impact and even benefits the soil environment is maize/soybean relay intercropping. This intercropping technique creates interspecific nitrogen competition between maize and soybean. Therefore, soybean nitrogen fixation is increased, which helps maintain soil fertility.[10] Increased soybean nitrogen fixation also means that the nitrogen in the soil is used more efficiently, and fewer external inputs of nitrogen (e.g., in the form of fertilizers) will be required.[11] In the long term, this can reduce the environmental impact of external chemicals added in soil, which will be discussed below. These two examples of intercropping demonstrate that by substituting monoculture with wisely planned cropping techniques, continuous cropping can be achieved with alleviated environmental harm, suggesting that agricultural productivity and environmental responsibility are not necessarily mutually exclusive.

2.2 External chemicals applied to promote crop development

2.2.1 Fertilizers

Fertilizers are widely applied in order to increase crop productivity. A global meta-analysis of 2,496 observations found that fertilizer application improved crop yield by 30.9% and nutritional quality by 11.9% on average.[12] Fertilizers achieve such effects primarily by supplying essential nutrients in plant-available forms, increasing nutrient availability in the soil, and facilitating nutrient uptake by crops. However, when fertilizers are applied to soils by broadcast, a considerable proportion of nutrients is not applied in easily accessible positions and is therefore hard for crops to take up.[13] For example, according to a global meta-analysis of 3,586 observations, the average nitrogen fertilizer use efficiency was only 39%.[14] This means a large proportion of nutrients fail to be used by the crops, some of which can potentially be washed away into waterbodies, causing water eutrophication and disrupting aquatic ecosystems. Eutrophication boosts the growth of algae in coastal waters. When those algae die, they create an increased accumulation of organic matter, which requires increased microbial activity and, therefore, increased microbial respiration to decompose. In the long term, increased oxygen consumption decreases the levels of dissolved oxygen in water, compressing available habitats, reducing secondary production, and altering the flow of energy through the food web.[15]

The ethical challenge raised by fertilization is how to reduce its impact on other species and ecosystems without sacrificing crop yield, which must continue to rise to feed the growing global population. A potential win-win solution is to shift to controlled-release

fertilizers. We will investigate how controlled-release fertilizers maintain or even increase crop yield with reduced environmental impact using controlled-release nitrogen fertilizers (CRN) as an example. CRN slows or controls nutrient release by coating readily available nitrogen fertilizer or by chemical modification. This means a greater proportion of applied nitrogen is taken up by crops and the amount of nitrogen lost to the environment is reduced. A global meta-analysis reveals that CRN can reduce overall nitrogen loss by 32.6% to 49.1%,[16] which can help reduce the occurrence of eutrophication. This is because CRN facilitates nitrogen uptake by providing better synchronization of crop nitrogen demand with nitrogen supply, eventually increasing the total amount of nitrogen uptake.[17] Besides eutrophication, CRN can also reduce other kinds of environmental pollution associated with fertilization, such as greenhouse gas emissions.[16] Meanwhile, there is evidence that CRN not only does not sacrifice crop yield but can even promote the productivity of various crops compared with traditional nitrogen fertilizers. This is because CRN affects how efficiently nitrogen is used for plant growth, particularly in photosynthesis. Experiments investigating the effects of CRN on sunflower growth proved that when applied at an appropriate rate, CRN can result in a 76.10% higher proportion of photosynthetic nitrogen than traditional nitrogen fertilizers applied at the equivalent nitrogen levels. Research has illustrated the mechanism of this difference: under conventional treatments, leaf nitrogen is often redistributed to support seed protein accumulation rather than photosynthesis in the later stages of crop growth. CRN, however, is able to maintain a high level of photosynthetic nitrogen content in the leaves during the later growth stages. Eventually, an increase in the amount of nitrogen used for photosynthesis results in better plant growth and higher yield.[17]

Another possible and more radical solution to the environmental impact of fertilizers is to shift to other methods for nutrient supply that require no or little external chemical input. While fertilizers promote plant growth by adding external nutrients to the environment, there is a considerable amount of nutrients that already exist in the environment in forms that are hard for crops to access (e.g., clays) and therefore fail to be used by crops. If those nutrients can be made available to crops, they can serve as an eco-friendlier substitute for nutrients from fertilizers. For example, research has shown that exopolysaccharides (EPS) from *Rhizobium tropici* can improve the bioavailability of soil legacy phosphorus in soils with substantial phosphorus accumulation. This type of EPS achieves this effect with its functional groups such as carboxyls and phenols that can suppress phosphate crystallization and increase the solubility of phosphate. Additionally, this type of EPS is nontoxic and biodegradable, making this method superior to conventional approaches that require inputs of synthetic chemicals.[18] From an ethical perspective, this approach is more radical and conceptually more compatible with the goal of minimizing human intervention in nature: rather than introducing external chemicals to directly alter the environment, it uses biological processes to make nutrients already present in the soil more available to crops.

Overall, fertilizers play an important role in crop development by providing nutrients that crops need for growth. However, the environmental consequences of conventional fertilizer use demonstrate the need to develop alternative nutrient-management strategies that can maintain adequate nutrient supply with reduced environmental impacts and, ultimately, a conceptual shift to strategies that minimize human intervention in natural systems. While the above-mentioned solutions have demonstrated potential effectiveness, their

practical application is not straightforward. For example, due to factors such as the high price of CRN, the amount of CRN applied in China only accounts for 1% of the total N fertilizer application.[16] Similarly, according to a 2025 review, few EPS have been industrially developed and commercially applied due to the low production yield and high costs.[19] It is understandable that stakeholders are currently more willing to stick to conventional methods that are more familiar and less expensive. Large-scale adoption of such new approaches therefore requires further scientific and technological innovations in related fields and potential policy support to encourage the transition from traditional farming practices to more novel and sustainable ones.

2.2.2 Pesticides

Besides fertilizers, pesticides are a major source of external chemical inputs into the environment involved in agricultural production. Pesticides are introduced to protect crops from pests and maintain yields. However, in practice, pesticides not only kill the target pests but also affect non-target organisms. A global meta-analysis of 20,212 effect sizes from 1,705 experimental studies found that pesticides negatively affect the growth, reproduction, behavior, and other physiological biomarkers of non-target plants, animals, and microorganisms.[20] In this way, pesticides disrupt ecosystems in the long term. Such effects are well documented in soil ecosystems. A meta-analysis of 54 studies and 294 observations found that pesticides decreased the abundance and diversity of soil fauna communities overall and that this effect was stronger on diversity than abundance.[21] Besides terrestrial ecosystems, aquatic ones can also be disrupted by pesticides when they enter waterbodies. Research has shown that increasing pesticide contamination results in decreasing regional aquatic biodiversity. Even at pesticide concentrations equal to the legally accepted regulatory threshold levels, which are oftentimes exceeded, aquatic macroinvertebrate family richness is reduced significantly.[22] Some of the ecological consequences of pesticides can reinforce themselves in the long term. For example, while pesticides kill pests, they also adversely affect non-target organisms, including predators that are beneficial to the ecosystem. As a result, outbreaks of secondary pests are expected, which may in turn require additional pesticide inputs. Even if pesticides affected only their target pests, however, extensive pesticide use could still select for pesticide-resistant individuals within pest populations, resulting in the evolution of pesticide resistance and reducing pesticide effectiveness, potentially leading to increased pesticide applications in order to maintain yield.[23] Thus, pesticides are unsustainable not only because of their direct chemical impacts on the environment but also because of the "vicious circle" they produce in the environment.

Biocontrol provides a sustainable alternative to pesticides. Instead of introducing external synthetic pesticides, biocontrol controls pest abundance by using botanical pesticides (insecticidal compounds or powder extracted from plants known to have pesticidal properties), introducing or releasing the natural enemies of pests, etc.[24,25] A meta-analysis of 99 studies across 31 crops found that, compared to synthetic pesticides, biocontrol resulted in comparable pest abundance and crop yield, while natural enemy abundance was 43% greater.[24] This demonstrates that biocontrol is able to achieve the same level of pest control as conventional pesticides while producing less disturbance to the ecosystem. More importantly, biocontrol is conceptually superior because it achieves pest control by working with natural ecological processes rather than relying on external chemical input. This method is a promising step to a more ethical future of

agricultural production because it acknowledges, respects, and wisely makes use of natural processes to reduce the need for human intervention. In practice, however, conventional chemical control remains, and is expected to remain, the dominant method of pest control due to technical, economic, and attitudinal barriers to the adoption of biocontrol. A broader transition toward biological pest control therefore requires technical advancements that make biocontrol more affordable and easier to implement, as well as greater public awareness and acceptance to encourage more farmers and agricultural stakeholders to adopt it.[25]

Scientists have also been developing another possible alternative approach to conventional pesticides using pro-natural products, which degrade into naturally occurring products after doing their job rather than leaving persistent residues. In this way, the environmental impact of pesticide inputs can potentially be reduced while retaining their intended functions. For example, scientists developed an ester-based pro-natural product that showed strong antifungal activity against diseases affecting rapeseed, rice, and apples. Meanwhile, after entering the soil, this product degrades rapidly into naturally occurring substances, with a half-life of roughly 20–31 hours. This example demonstrates that environmental protection does not require eliminating all kinds of external chemical inputs. By mitigating the long-term ecological consequences of such inputs, they can align with our sustainability goals while still performing their intended functions.[26]

3. Ethical and Sustainability Challenges Associated with Animal Husbandry

Besides crop production, modern agriculture also involves the breeding, management, and exploitation of animals. Animal husbandry affects the environment through various mechanisms. Livestock emit greenhouse gases by exhaling CO_2, and their manure release additional greenhouse gases such as CO_2, CH_4 and N_2O, resulting in climate change and air pollution.[27,28] Livestock emit 73% of all agricultural nitrogen to water bodies, contributing to eutrophication.[28] Additionally, livestock activities such as grazing can lead to topsoil compaction, which is shown to raise N_2O emissions.[29]

When humans start to not only exploit natural resources like land but also animals for their own benefit, the consequences extend beyond direct environmental impact to more severe ethical concerns. The most obvious ethical problem regarding animals is animal welfare, which is concerned with both the physical and emotional well-being of animals.[30] In fact, animal welfare is not only an ethical concern. From a practical perspective, decreases in reproductive efficiency are associated with increased animal stress, which means failure to protect animal welfare will potentially compromise the productivity and sustainability of agricultural production.[30] One of the most important factors that affect animal welfare is the environment in which animals are kept. Economic profits encourage producers to increase stocking density, which poses a serious threat to animal welfare. For example, research has shown that the well-being of broiler chickens, one of the most prevalent farm animals, is threatened at high stocking density. Higher stocking density is associated with increased daily mortality, increased incidence of diseases such as leg disorders, disturbed resting behavior, decreased locomotion and ground pecking, increased psychological stress, etc.[31,32] Even when stocking density does not exceed the limit, the quality of the housing environment can still limit the health and welfare of broilers.

Humans not only manipulate how animals live but also how they reproduce, creating a more complex and serious ethical issue. Humans have long been selecting specific

traits in animals to achieve increased production, a classic example being the selection for increased milk production. This selection is facilitated by artificial insemination, where semen is collected from a male animal and manually introduced into the female reproductive tract without natural mating, allowing individuals with preferred traits to have more offspring. For example, in U.S. Holstein cattle, genetic selection increased annual milk yield by 79% during the 40 years of 1964–2004.[33] Meanwhile, since this selection approach is mostly based on traits that can be easily observed and recorded, the possibility of unexpected genetic defects increases when a population contains a large number of individuals that are descendants of the same individual, resulting in more frequent inbreeding. A study focusing on livestock species, including cows, pigs, and sheep, reveals that one of the traits that are not easily observed and have therefore been neglected by this selection approach is fertility, which worsens the consequences of artificial selection because fertility is the single most important trait in all livestock species.[30] For the U.S. Holstein cattle case, the increase in milk yield is accompanied by an approximately 30-day increase in the time required to achieve successful conception.[33]

Overall, the ethical concerns related to animals in agricultural production are generally caused by humans manipulating animals as populations that can be artificially managed and genetically reshaped for short-term benefits. While the international community has noticed this challenge and related regulations have been implemented, many of these measures remain insufficient to fully address the underlying ethical concerns. For example, the European Union (EU) has implemented animal welfare legislation to ensure good housing and food conditions for animals kept for farm purposes, to protect animals from suffering during transport, to standardize the conditions at the time of stunning and slaughter, etc.[34] However, such legislation mostly focuses on resources and management (e.g., average area of land for an animal, amount of nutrients) and often neglects animal-based indicators that directly reflect animal welfare (e.g., frequency of lameness, injuries, disease, and mortality).[35] While it is understandable that current legislation emphasizes the application of resources, which is easier than directly assessing animal well-being, this focus does reveal that current legislation is conceptually limited because it still originates from a human-centered perspective and fails to treat animals as individuals that can experience both physical and emotional suffering.

4. Conclusion

Modern agriculture has enabled enormous increases in food production, but these benefits have come at environmental and ethical costs. This paper discusses the ethical challenges associated with modern agricultural production and explores potential alternative practices. From continuous cropping, fertilizers, and pesticides that disrupt the soil environment and ecosystems, to animal husbandry that neglects the protection of animal welfare and the potential long-term consequences of genetic selection, agriculture has produced significant negative consequences for the environment and the well-being of non-human species, disrupting complex natural processes. Essentially, such concerns are raised because humans are approaching and exploiting nature as a resource instead of a habitat. Consequently, humans and nature become opposite parties rather than a harmonious whole, and real shift to more sustainable and ethical agriculture would not be possible if this perspective remained unchanged. This is why, although alternative agricultural practices have been designed and implemented to potentially reduce the impact on the environment and non-human species, many of

them are conceptually limited with their human-centered perspective. To make a true difference, it would be helpful to develop a renewed human-nature relationship that can be described as "Symbiocene", a concept that refers to humans reintegrating themselves into nature. This involves a shift in perspective from economy-centered thinking to planet-centered thinking when carrying out production decisions that involve nature. It also urges humans as a species to stay humble and be aware that what we do to nature could result in unforeseeable consequences that would be hard to reverse once we notice them in the future. Practically, "Symbiocene" also requires structural and institutional changes, such as technological advancements to design better practices, policies to encourage the adoption of these practices, etc.[36] Ultimately, the path to a more sustainable and ethical agriculture involves a combination of social, economic, political, and ideological challenges. Despite the complex challenges, a shift away from a purely pragmatic, human-centered approach toward one that recognizes and respects animal well-being and ecosystem integrity would be a promising start that would help us find the right direction.

5. References

[1]Gignoux, C. R., et al. (2011). Rapid, global demographic expansions after the origins of agriculture. *Proceedings of the National Academy of Sciences, 108*(15), 6044–6049. https://doi.org/10.1073/pnas.0914274108

[2]Foley, J. A., et al. (2005). Global consequences of land use. *Science, 309*(5734), 570–574. https://doi.org/10.1126/science.1111772

[3]United Nations. (2024). *State of world population 2024*. Stylus Publishing.

[4]Chen, Y., Du, J., et al. (2022). Evolutions and managements of soil microbial community structure drove by continuous cropping. *Frontiers in Microbiology, 13*. https://doi.org/10.3389/fmicb.2022.839494

[5]Bai, Y., et al. (2019). Soil acidification in continuously cropped tobacco alters bacterial community structure and diversity via the accumulation of phenolic acids. *Scientific Reports, 9*(1). https://doi.org/10.1038/s41598-019-48611-5

[6]Chen, Y., Yang, L., et al. (2023). Autotoxins in continuous tobacco cropping soils and their management. *Frontiers in Plant Science, 14*. https://doi.org/10.3389/fpls.2023.1106033

[7]Yu, J., et al. (2022). The cropping obstacle of garlic was associated with changes in soil physicochemical properties, enzymatic activities and bacterial and fungal communities. *Frontiers in Microbiology, 13*. https://doi.org/10.3389/fmicb.2022.828196

[8]Du, L., et al. (2017). Effects of garlic/cucumber relay intercropping on soil enzyme activities and the microbial environment in continuous cropping. *HortScience, 52*(1), 78–84. https://doi.org/10.21273/hortsci11442-16

[9]Xiao, X., et al. (2019). A green garlic (*Allium sativum* L.) based intercropping system reduces the strain of continuous monocropping in cucumber (*Cucumis sativus* L.) by adjusting the micro-ecological environment of soil. *PeerJ, 7*, e7267. https://doi.org/10.7717/peerj.7267

[10]Du, Q., et al. (2020). Relay-intercropping soybean with maize maintains soil fertility and increases nitrogen recovery efficiency by reducing nitrogen input. *The Crop Journal, 8*(1), 140–152. https://doi.org/10.1016/j.cj.2019.06.010

[11]Xu, Z., et al. (2020). Intercropping maize and soybean increases efficiency of land and fertilizer nitrogen use; a meta-analysis. *Field Crops Research, 246*, 107661. https://doi.org/10.1016/j.fcr.2019.107661

[12]Ishfaq, M., et al. (2023). Improvement of nutritional quality of food crops with fertilizer: A global meta-analysis. *Agronomy for Sustainable Development, 43*(6). https://doi.org/10.1007/s13593-023-00923-7

[13]Nkebiwe, P. M., et al. (2016). Fertilizer placement to improve crop nutrient acquisition and yield: A review and meta-analysis. *Field Crops Research, 196*, 389–401. https://doi.org/10.1016/j.fcr.2016.07.018
[14]Yu, X., et al. (2022). Global meta-analysis of nitrogen fertilizer use efficiency in rice, wheat and maize. *Agriculture, Ecosystems & Environment, 338*, 108089. https://doi.org/10.1016/j.agee.2022.108089
[15]Diaz, R. J., & Rosenberg, R. (2008). Spreading dead zones and consequences for marine ecosystems. *Science, 321*(5891), 926–929. https://doi.org/10.1126/science.1156401
[16]Zhang, Y., et al. (2024). Substituting readily available nitrogen fertilizer with controlled-release nitrogen fertilizer improves crop yield and nitrogen uptake while mitigating environmental risks: A global meta-analysis. *Field Crops Research, 306*, 109221. https://doi.org/10.1016/j.fcr.2023.109221
[17]Ren, W., et al. (2025). Controlled-release fertilizer affects leaf nitrogen allocation and photosynthesis to improve nitrogen use efficiency and yield in the sunflower field. *Frontiers in Plant Science, 16*. https://doi.org/10.3389/fpls.2025.1622766
[18]Xie, X., et al. (2026). Exopolysaccharides from *Rhizobium tropici* enhance the solubility of soil legacy phosphorus. *ACS Earth and Space Chemistry, 10*(2), 592–602. https://doi.org/10.1021/acsearthspacechem.5c00377
[19]Liu, L., et al. (2025). Advances and challenges in bioproduction of microbial exopolysaccharides: Synthesis mechanisms, engineering strategies, and future perspectives. *Carbohydrate Polymers, 367*, 124010. https://doi.org/10.1016/j.carbpol.2025.124010
[20]Wan, N.-F., et al. (2025). Pesticides have negative effects on non-target organisms. *Nature Communications, 16*(1). https://doi.org/10.1038/s41467-025-56732-x
[21]Beaumelle, L., et al. (2023). Pesticide effects on soil fauna communities: A meta-analysis. *Journal of Applied Ecology, 60*(7). https://doi.org/10.1111/1365-2664.14437
[22]Stehle, S., & Schulz, R. (2015). Agricultural insecticides threaten surface waters at the global scale. *Proceedings of the National Academy of Sciences, 112*(18), 5750–5755. https://doi.org/10.1073/pnas.1500232112
[23]Pimentel, D. (2005). Environmental and economic costs of the application of pesticides primarily in the United States. *Environment, Development and Sustainability, 7*(2), 229–252. https://doi.org/10.1007/s10668-005-7314-2
[24]Ratto, F., et al. (2022). Biological control interventions reduce pest abundance and crop damage while maintaining natural enemies in sub-Saharan Africa: A meta-analysis. *Proceedings of the Royal Society B, 289*(1988). https://doi.org/10.1098/rspb.2022.1695
[25]Bale, J. S., et al. (2007). Biological control and sustainable food production. *Philosophical Transactions of the Royal Society B, 363*(1492), 761–776. https://doi.org/10.1098/rstb.2007.2182
[26]Lei, P., et al. (2026). A pro-natural product design strategy affords ester-based green fungicide candidates. *Science Advances, 12*(35). https://doi.org/10.1126/sciadv.aej0610
[27]Philippe, F. X., & Nicks, B. (2015). Review on greenhouse gas emissions from pig houses: Production of carbon dioxide, methane and nitrous oxide by animals and manure. *Agriculture, Ecosystems & Environment, 199*, 10–25. https://doi.org/10.1016/j.agee.2014.08.015
[28]Uwizeye, A., et al. (2020). Nitrogen emissions along global livestock supply chains. *Nature Food, 1*(7), 437–446. https://doi.org/10.1038/s43016-020-0113-y
[29]Pulido-Moncada, M., et al. (2022). Soil compaction raises nitrous oxide emissions in managed agroecosystems. A review. *Agronomy for Sustainable Development, 42*(3). https://doi.org/10.1007/s13593-022-00773-9

[30]van Marle-Köster, E., & Visser, C. (2021). Unintended consequences of selection for increased production on the health and welfare of livestock. *Archives Animal Breeding, 64*(1), 177–185. https://doi.org/10.5194/aab-64-177-2021

[31]Hall, A. L. (2001). The effect of stocking density on the welfare and behaviour of broiler chickens reared commercially. *Animal Welfare, 10*(1), 23–40. https://doi.org/10.1017/s096272860002323x

[32]Estevez, I. (2007). Density allowances for broilers: Where to set the limits? *Poultry Science, 86*(6), 1265–1272. https://doi.org/10.1093/ps/86.6.1265

[33]Ma, L., et al. (2019). Genome changes due to artificial selection in U.S. Holstein cattle. *BMC Genomics, 20*(1). https://doi.org/10.1186/s12864-019-5459-x

[34]European Commission. (2023). *EU animal welfare legislation*. Food Safety. https://food.ec.europa.eu/animals/animal-welfare/eu-animal-welfare-legislation_en

[35]Dusel, S., & Wieck, C. (2023). Evidence gaps hinder animal welfare progress in the European Union. *Nature Food, 4*(5), 348–349. https://doi.org/10.1038/s43016-023-00743-8

[36]Heger, T., & Szasz, O. (2026). The Symbiocene as a transformative vision. *People and Nature*. https://doi.org/10.1002/pan3.70415

Domestic Regulation and Global Responsibility: The Case of Desflurane

William Neice[1*], Anish Peri[2*], Joel Pae[3*], William Jiang[4*]
[1]1 Lake Oswego High School, OR 97034, [2]Eden Prairie High School, MN 55346, [3]Korea International School, Gyeonggi-do 13543, South Korea, [4]The Harker School, CA 95129

Abstract

Greenhouse gas emissions from halogenated anesthetics declined by around 27% between 2014 and 2023, cited as progress towards minimizing anesthetic emissions. However, when separated by income, greenhouse gas impact specifically due to desflurane has declined by 52% among high-income nations, increased by 151% among upper-middle-income nations, and increased by 2,281% within low and lower-middle-income countries. These trends have been driven disproportionately by desflurane, an anesthetic that offers little clinical benefit over sevoflurane and a global warming potential roughly 2,590 times that of carbon dioxide. The UK and EU have banned desflurane, but it is estimated that full compliance would reduce global impact by only 14%, since manufacturing and export remain largely untouched. This paper asks whether a domestic phase-out leaving production and export untouched discharges a nation's responsibility for harmful products it continues to send abroad. Considering arguments from climate justice and pollution haven theory, it argues that desflurane use is an unnecessary luxury and bans on consumption merely relocate its usage rather than eliminate it. Historical precedents suggest that unilateral restrictions on globally traded hazards are often ineffective, whereas broader international agreements address issues more completely. The paper also considers competing explanations for increasing desflurane: the demonstration effect, market incentives, and clinician awareness. It also examines whether the GWP calculation overstates desflurane's climate significance, and how figures involved in the debate have financial relationships with the manufacturer. Rather than relying on either side, it supports investment in anesthetic gas capture and recycling technology that reduces damage at the point of use. Regional bans reduce emissions where implemented, but do not necessarily address continued production and export of the drug, growing use elsewhere, or economic incentives supporting that use, making them necessary but insufficient without a coordinated global response.

1. Introduction

Global greenhouse gas emissions from the halogenated anesthetic agents used in general anesthesia, namely sevoflurane, desflurane, isoflurane, and halothane, dropped by roughly 27% between 2014 and 2023. This figure has been widely celebrated as evidence of the anesthesia community's progress against climate change [1]. During that same decade, however, the greenhouse gas impact from desflurane specifically decreased by 52% in high-income countries, while it increased by 151% in upper-middle income countries and by a staggering 2,281% in lower-middle- and low-income countries [1]. This growth is primarily concentrated in a handful of middle-income markets, such as China, Indonesia, Malaysia, Russia, Vietnam, and South Africa [1]. Although these findings show diverging patterns in national income-groups, they do not establish a correlation between declining use in specific regions and increasing use elsewhere. This matters because desflurane is less potent than sevoflurane, meaning the excess dosage necessary to achieve the same anesthetic effect causes its greenhouse gas impact to be 40 to 50 times higher in clinical

practice [1] — in spite of offering little clinical advantage beyond a marginally faster wake-up time. In the United States alone, desflurane contributed just 2.4% of volatile anesthetic usage in 2023, yet was responsible for 58.9% of the associated CO2 emissions from these compounds. Eliminating desflurane entirely, without any other changes, would cut total US anesthetic gas emissions by 57.6% [2] as a result of this severe disproportion.

Along with these historical trends, several jurisdictions have introduced desflurane use restrictions. For instance, desflurane use has been restricted in the UK since 2024 and in the EU beginning January 1, 2026, except in cases where its use is deemed imperative [1], [2], [3]. These policy shifts were driven by professional advocacy from bodies including the World Federation of Societies of Anesthesiologists, the European Society of Anaesthesiology and Intensive Care, and the American Society of Anesthesiologists [1]. Notably, the researchers compiling data on this trend project that an EU and UK phase-out, even if fully realized, would reduce desflurane's global impact by only 14% [1].

The anesthesia community is divided on how to interpret this relocation. Researchers analyzing US purchasing data believe that continued desflurane use by high-income countries may in fact promote the drug's perceived legitimacy elsewhere. Accordingly, if respected, resource-rich health systems such as the United States phase out desflurane's use, it could curb usage worldwide by removing the example justifying its continued adoption [2]. However, the dataset's authors point to a simpler culprit: market forces are driving the relocation. As desflurane sales rise across several middle-income countries — such as China, Indonesia, Malaysia, Russia, Vietnam, and South Africa — they speculate that "more competitive pricing or marketing of desflurane that specifically targets these countries" may be a contributing factor [1]. This paper takes the latter view. It argues that the relocation described above is better explained by an unaddressed incentive structure — manufacturing, marketing, and cost — rather than by the presence of clinical prestige or lack thereof, and that this distinction has real consequences for what an ethically adequate response looks like. If demand is being actively cultivated in markets with weaker regulatory capacity, then a domestic phase-out that leaves manufacturing and export untouched addresses the symptom in one jurisdiction while leaving the underlying cause fully intact.

This paper presents a literature-based ethical and policy analysis rather than an original empirical study. It draws on peer-reviewed empirical studies that examine global anesthetic emissions, clinical comparisons of volatile anesthetic gases such as those mentioned previously, and anesthetic gas capture technologies; scholarly work on environmental ethics and economics, including distributive justice, luxury emissions, and pollution haven theory; and primary legal and policy documents, including the EU F-gas Regulation, the Rio Declaration, the United Nations Framework Convention on Climate Change, the Rotterdam Convention, and the Montreal Protocol. Peer-reviewed commentaries and responses to the debate over desflurane's climate impact are evaluated as arguments rather than treated as independent empirical findings, while journalistic accounts and policy commentary provide historical context required when discussing regulation precedents. Sources were identified through Google Scholar and Stony Brook University Libraries' SEARCH discovery system using combinations of keywords including "desflurane", "volatile anesthetic emissions", "global warming potential", "anesthetic gas capture", "climate justice", "luxury emissions", "pollution haven

theory", and "international environmental regulation."

The main question this paper asks is whether or not it is ethical for wealthy nations to address the global issue of desflurane emissions by banning it in their own jurisdictions, while they continue to allow the manufacture and export abroad of the same drug to countries potentially lacking the same regulatory capacity to enforce similar bans. Section 2 will discuss the pattern of desflurane relocation in more detail, as well as the finding that per-capita impact varies significantly even among nations that are all high-income, preventing the formation of a simple narrative that wealthier countries are more virtuous in this regard. Section 3 will then move towards looking at the issue through a distributive justice framework. We will investigate whether the costs and environmental responsibilities of addressing these desflurane emissions are distributed equally among nations with different amounts of wealth or levels of regulatory capacity. Next, Section 4 examines whether wealthy nations should still bear responsibility for continued harm if they allow the manufacture and export of this drug while restricting domestic usage, drawing on precedents surrounding other globally traded yet selectively banned substances. Section 5 addresses the demonstration-effect hypothesis as the strongest available counterargument. Finally, Section 6 argues that a more ethical pathway forwards in addressing such emissions is by focusing on anesthetic gas recapture and recycling technology, rather than shifting regulatory burdens between different countries. However, until the economic incentives that sustain the global desflurane demand are directly addressed, regional phase-outs remain a necessary — but insufficient — response.

2. Diverging Trends in Desflurane Use

The ethical concern with unilateral phase-outs becomes clearer when we take a closer look at the 27% global decline: that figure masks substantial differences in how anesthetic emissions have changed across various nations. The data in this section distinguishes between the overall impact of halogenated anesthetic agents and the impact specifically attributable to desflurane use. The two measurements illustrate related but quantitatively different trends. Talbot, Holländer, and Bentzer based their estimate off IQVIA MIDAS sales data that covered approximately 80% of the global population, or 91 countries, between 2014 and 2023 [1]. Their estimate of a decline in the global greenhouse gas impact of halogenated anesthetics, from 2,754 to 2,005 kilotons of CO2-equivalent over the decade, is accurate at the global level. However, this decline depends on income tiers. High-income nations made up 92% of the total impact mirroring the necessity of a similar 'global agreement' in 2014 and still 71% in 2023, meaning that even after a decade of reform, wealthy countries remained the highest emitters by far. The change here does not concern the countries that emit the most, but rather the trends each income tier follows. High-income countries saw a 44% decrease in impact over this time period. Upper-middle-income countries saw a 129% rise, while low-income and lower-middle-income countries saw a 372% rise [1]. Desflurane displays an even more defined version of this same pattern, and in fact drives nearly all of the divergence already noted above [1].

It is important to grasp just how disproportionate desflurane's impact is. Desflurane's global warming potential over a 100-year period (GWP100) is roughly 2,590 times that of carbon dioxide, whereas sevoflurane's GWP100 is a much lower 144–195, depending on the source [1], [2].

Furthermore, GWP alone understates the drug's impact, since desflurane is also less potent, meaning a larger dosage is necessary in achieving the same anesthetic depth. Simulations that combined potency and warming potential found that a single standard two-hour anesthesia session using desflurane rather than sevoflurane released approximately 45.5 times the emissions [1]. As noted above, the same disproportion appears in US national data [2].

Another analysis, based on a different but overlapping dataset over the same time period of 2018–2023, covered 41 countries and supported the aforementioned divergence while also providing more information about where geographically the increase is occurring. 57% of the total greenhouse gas impact from these agents in 2023 came from the United States, European Union, and Japan combined [1]. Separately, the 41-country analysis found that China's emissions increased by 130% over that time period to make it the third largest national emitter by absolute terms in the dataset with 277,882 tonnes of CO2-equivalent. The United States and the European Union came before it at 443,312 tonnes and 410,726 tonnes, respectively [4]. However, when looking at per-capita emissions, Japan and South Korea had the highest rates measured at approximately 2.5 kg CO2-equivalent per inhabitant — five to twenty-five times the per-capita rate reported for many Western-European nations moving furthest towards desflurane phase-out.

It is difficult to form neat groups consisting of countries currently responding to the problem and those that are not. As a general rule, high-income countries have been moving towards reducing desflurane use, while in many middle-income countries usage has been increasing. Nonetheless, substantial variation remains within each group. Japan, Spain, France, and the United States have per-capita emissions from halogenated anesthetics that are nearly ten times those of Sweden and the Netherlands, even though all the countries are

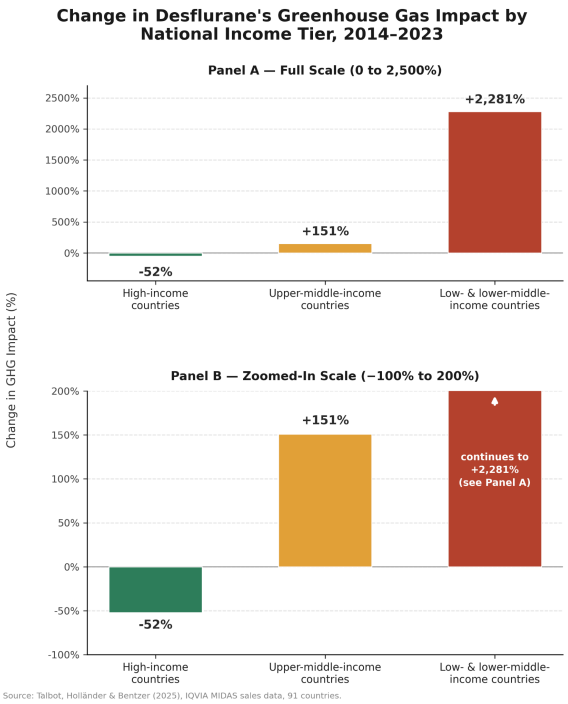


Figure 1: Change in desflurane's greenhouse gas impact by national income tier, 2014–2023 (Talbot et al., 2025)

within the high-income group [1]. Differences in how desflurane is actually used explain some, but not all, of this discrepancy: even after converting desflurane to an equivalent amount of sevoflurane, Sweden and the Netherlands's emissions remain 2-10 times lower than the highest-emitting countries of their income level [1]. Within the European Union, differences between individual countries are also difficult to explain in terms of regulation alone. The 41-country analysis saw the ratio between the highest- and lowest-emitting European countries fall from a 43-fold difference in 2018 (2.6 kg CO_2e per capita in France vs. 0.1 kg in Montenegro) to a 17-fold difference by 2023 (1.7 kg CO_2e per capita in the Czech Republic vs. 0.1 kg in Poland) [4]. Although the gap has been massively reduced, a 17-fold gap between countries bound by the same EU regulations is

still enormous, and suggests clinical practice rather than regulation as the remaining explanation. Some potential contributors to these differences in emissions are fresh flow rates, the usage of regional versus intravenous anesthesia, and the use of automated low-flow systems. These cannot be explained by national wealth or regulation alone, so this complicates the idea that wealthier nations are more accountable. Regulations can change the compounds that hospitals are allowed to use, but not how they are administered, so some nations have made significant reductions in emissions while others still have high per-capita impacts.

Responses to this issue have so far been jurisdictional rather than global. In the UK, restrictions on desflurane use began in 2024, while the EU's F-gas Regulation 2024/573 prohibits its use beginning January 1, 2026, unless "such use is strictly required and no other anaesthetic can be used on medical grounds" [1], [3]. The authors of the study containing the largest global dataset modeled what would happen if the regulation were fully implemented, and estimated that while the EU and UK could reduce their own contribution to desflurane's environmental impact by as much as 69%, the reduction in desflurane's global impact would be only 14%. This is due to the EU and UK together remaining a non-majority of worldwide use, despite accounting for a substantial share [1]. The authors, based on this finding, concluded that "a global agreement to decommission desflurane is warranted" [1] since even a complete phase-out in certain regions would address only a limited portion of the global problem.

The data findings alone are insufficient in explaining why desflurane use is rising where it is. The primary study's authors do not provide any definitive findings, but the pricing and marketing hypothesis already noted above is what this paper's argument builds from [1].

3. A Distributive Justice Framework for Evaluating Unilateral Climate Action

Before the desflurane relocation discussed in Section 2 can be utilized in an ethical argument, it is essential to establish the standards based upon which a unilateral climate action should be judged. In this case, the principle of common but differentiated responsibilities (CBDR) and the distinction between subsistence and luxury emissions made by Henry Shue appear to be most effective. Both were established for different problems than this paper analyzes but if utilized carefully, they expose the fundamental flaws of a domestic-only desflurane elimination policy.

The CBDR has formed the basis of international environmental regulation since 1992 after appearing directly in the Rio Declaration's Principle 7, stating that states have an obligation to ensure environmental protection, but one which falls disproportionately upon each state according to both its past contribution and present capabilities [10]. The United Nations Framework Convention on Climate Change (UNFCCC), signed concurrently in 1992, incorporates this idea more directly through its Article 3.1, holding that states are to protect "the climate system on the basis of equity and in accordance with their common but differentiated responsibilities and respective capabilities." [11] This specifically charges developed nations with the responsibility to take immediate action and establish harsher policies than those of developing countries, which is exemplified by the prohibition of desflurane by the EU and UK. However, there exists an additional dimension of this issue that the CBDR has not been asked to address: can a wealthy nation fulfill its obligation to reduce global greenhouse gas emissions when it reduces only its domestic emissions despite retaining significant production or distribution capacities? Such a reading of "capacity" should not be interpreted

this narrowly. Capacity, in the wording of the Rio Declaration, refers to a nation's ability to control technologies and financial means to solve environmental issues [10]. This extends to the regulation of exports by its national companies, not merely to what its hospitals are allowed to purchase.

The distinction made by Henry Shue between subsistence emissions and luxury emissions further clarifies the matter. Shue argued that greenhouse gases emitted for purposes of sustaining one's basic survival needs are morally distinguishable from those emitted through discretionary consumption choices, and climate policies failing to take such a distinction into account risk asking the disadvantaged to make sacrifices to their basic necessities in favor of preserving comforts of the rich [12]. Applied to anesthesia, this distinction initially seems to contradict this paper's claims: general anesthesia is not a luxury. It is essential medical care and a framework that treats it as the equivalent of private jet travel, for example, would be unfair to healthcare systems that depend on it. Yet the goal of this paper is not to single out the environmental footprint of general anesthesia. Instead, it concerns the differing greenhouse gas footprint incurred from preferring desflurane over sevoflurane for modestly faster emergence times [1]. This difference is a closer fit to what Shue would call a luxury emission: not the underlying medical necessity but the convenience associated with the slight benefits of desflurane, coming at a 45.5-times higher environmental cost. This increment means that Shue's distinction does not exempt desflurane from scrutiny; it just specifies which characteristics must be scrutinized.

This formulation also exposes an imbalance that CBDR alone does not account for. Shue's subsistence and luxury distinction was designed to protect the global poor from having their essential emissions limited by policies designed for the wealthy. However, nothing prevents this logic from being applied from the poor to the wealthy. Regardless of where it is being utilized, desflurane still provides the same luxury benefits. It does not matter if its use is practiced by a wealthy country's anesthesiologists or a lower-middle-income country's anesthesiologists. What differs is not the morality of the emissions, but which country faces a market barrier to making desflurane and which does not. A domestic phase-out that leaves manufacturing and export untouched does not eliminate this luxury increment; it simply relocates the freedom to benefit from it to different healthcare systems where the anesthetic's sale remains profitable. The primary dataset found that this is exactly what has happened, as desflurane's sale has risen dramatically in middle-income markets after a marketing push [1].

This is not a claim that blame is distributed evenly. Shue's distinction identifies the moral status of an emission by its purpose, not by the wealth of the country burning it; a marginal convenience remains a marginal convenience regardless of where it is chosen. Responsibility for that choice being available at all is a separate question, and on this question CBDR's graduated standard continues to apply in full: it is the manufacturer marketing desflurane into markets with weaker bargaining power, not the anesthesiologist working with the options a fragile supply chain has given them, who bears the greater share of blame for the emission's existence. Taken together, the two principles do not compete, but rather divide labor. Shue's framework tells us that the harm itself does not become more essential by relocating, while CBDR tells us who is responsible for undoing that relocation — and the answer to that question points back to the nation exporting the drug, not the hospital importing and using it.

There is a similar theory in trade-and-environment economics to demonstrate why this phenomenon could be expected [13]. The concept of the pollution haven theory posits that stricter regulations within a given territory do not lead to reduced levels of pollution, as this merely results in pollution relocating to territories with more lenient regulations — potentially increasing global pollution levels, assuming that the recipient territory has reduced production efficiencies. This theory is primarily designed to study manufacturing relocation; however, what has been observed in this case, as discussed in Section 2, represents precisely this mechanism in action, except as a demand-driven process rather than in manufacturing. Desflurane producers have not relocated their factories but the drug's usage has relocated, adhering to identical economic principles for the benefit of their business interests. Specifically, any enterprise operating under reduced demand due to increasing regulation within a specific geographic area will benefit commercially from finding a source of continued demand that remains unregulated. Furthermore, the authors of the primary dataset independently reached a similar conclusion regarding the drivers behind increasing desflurane usage within certain nations [1].

These three approaches combined form the standard for the rest of this paper. For a unilateral climate policy initiative to satisfy a wealthy country's differentiated responsibility to act upon global warming, it must ensure that no means remain for shifting environmental damages abroad. A domestic ban that leaves manufacturing, marketing, and exports untouched does not qualify under this definition, not due to a lack of moral intent behind the ban but rather because it confuses the concepts of reducing one's own environmental impact versus minimizing overall damage to the planet. Sections 4 and 5 test this standard against the precedents and the strongest counter-arguments.

4. Precedents for Alternative Selectively Banned Substances

Examining three distinct case studies highlights that desflurane, in fact, is not the first product restricted among wealthy nations for its own population due to health and environmental concerns, yet also actively exported by them. Agricultural pesticide exports may bear the closest resemblance to the case for desflurane, as it demonstrates the standing of a domestic ban operating alongside legal, continuous export. Leaded gasoline furthermore illustrates a national hazard's multi-decade phase-out amidst demonstration and market forces, while the Montreal protocol depicts a globally traded hazard addressed through binding, financed international contracts, mirroring the necessity of a similar "global agreement" for desflurane as called for by the primary dataset's authors [1].

4.1. Leaded Gasoline

Tetraethyl lead (TEL) was added to gasoline in the 1920s, and upon its recognition as a public health hazard in the 1970s, wealthy countries were prompt to enact legislative policies—Japan banned leaded gasoline (LG) in 1986, and the US eliminated its use in vehicles by the mid-1980s, followed by a full ban in 1996 [14]. Despite these sweeping measures, LG continued to exist for more than 30 years; by the early 2000s up to 117 nations were still actively utilizing it, with complete global elimination occurring not until July 2021, when Algeria depleted its final stockpile [15]. A common misconception is that this wide margin between prohibition and eradication of LG was not due to reforms in developed nations providing examples for remaining regions of the world. Rather, its removal was catalyzed by the United Nations Environment Programme, which ran a campaign lasting nearly two decades with

socioeconomic resources since 2002 to phase-out fuel usage via cooperation with individual governments and oil companies, a sustained international approach [14]. According to an official involved with the program at the time, however, the fact that wealthy countries expedited changes to prevent air pollution caused them to overlook the continued persistence of environmental harms from LG elsewhere in the world; this is particularly notable, since it appears neither pragmatic nor logical for such a separated focus on issues pertaining to climate change which affects the entire globe. Furthermore, the cause for perpetuated LG use throughout the early 21st century was not due to ignorance of environmental degradations, but instead attributable to the fact that TEL additive manufacturing industries paid national officials to purchase large stockpiles before its disappearance from the market, an effort that directly countered the initiative to reduce the hazard [14]. The case study of leaded gasoline is a historical instance that portrays market-forces mechanisms analogous to desflurane's current trajectory, as mentioned in Section 2. Thus, while this anesthetic's prominence is diminishing in the market of wealthy countries, its environmental harms can only truly be mitigated with the addressal of stakeholders actively working to preserve demand elsewhere.

4.2. Agricultural Pesticides

Agricultural pesticides offer another comparison to desflurane; whereas leaded gasoline represented a globally fragmented phase-out, the case for this product is a continuously operating export channel amidst domestic bans. Under US law, pesticides that are banned, restricted, or never registered for domestic use can be manufactured and exported as long as labeling and notification guidelines are satisfied. Because the requirements are quite minimal, 2013 data from the Environmental Protection Agency (EPA) showed pesticide manufacturing for export occurred in up to 23 US states [16]. Between 2015 - 2019, pesticides with neurotoxic organophosphates and carbamates approved for domestic use were also exported to 42 different countries, as found by the recent analysis [16]. While this wave of continued exports is already problematic on its own, the issue is aggravated by the fact that much of this chemical-grown produce is eventually imported back through food grown without adequate equipment or protective standards from workers in countries with weaker regulations for exposure management [16]. In effect, this renders a "circle of poison"—the paradoxical nature of which stems from how the same substance deemed too dangerous by a country returns to it unchecked. Similarly, despite desflurane posing an unacceptable risk for a country's own atmosphere, it remains entirely legal for that same nation's manufacturers to sell it in markets without the capabilities of weighing or resisting that risk, further compromising the possibility of domestic cycling.

The international response to this pattern is a counterexample of the argument outlined in this paper. The Rotterdam Convention, ratified by more than 160 countries, was adopted in 1998 and formally executed in 2004. Instead of banning export of hazardous chemicals or pesticides directly, it relies on prior informed consent procedures. Exporting countries need to notify and obtain the formal consent of the importing country for listed substances before shipment, providing adequate information of the substance's legal status for the receiving government to consider [17], [18]. While the limited premise of this convention is more pragmatic than a global ban, the "poison cycle" it implicitly permits obscures the deeper dilemma of evaluating whether the substance should be banned globally, as opposed to being restricted and permitted in select regions.

Despite the sheer magnitude of per-dose climate effects noted in Section 2, no consent requirement policies currently exist for exporting halogenated anesthetic agents; this absence is not because such a mechanism is inopportune, but rather attributable to the lack of regulatory attention for a pharmaceutical whose principal harm is atmospheric and aquatic instead of acute toxicity for humans, the premise framework outlined in Section 6.

4.3. The Montreal Protocol

While the case studies above demonstrate the limitations of uneven national responses to environmental harms, the Montreal Protocol provides a complementary example of how international cooperation can address issues more effectively, supporting the case for global agreement called for by the dataset's primary authors [1]. The Montreal Protocol was adopted in 1987 to eradicate chlorofluorocarbon and related chemicals, and by 2009 it was the only environmental treaty to achieve universal ratification by all 197 countries and eliminated approximately 98% of targeted substances [21]. Unlike the Rotterdam Convention and the uncoordinated phase-out of leaded gasoline as discussed above, the Protocol established different compliance timelines in its clauses, which minimized influence by market forces or moral suasion to give developing and developed countries enough time for them to transition [19]. It also established the Multilateral Fund, the first international financial mechanism to support developing countries with transition costs and technology transfer, providing a tangible economic support base rather than an unfunded mandate [19]. The protocol also crucially addressed the potential issue of production shifting to nations outside of those participating in the agreement by restricting trade of controlled substances with non-parties, making chemical production commercially disadvantageous for countries opting out [19]. Unlike the lacking enforcement and consent requirements of the precedents mentioned previously, the Montreal Protocol aligned economic incentives across countries, limiting the ability for environmental harms to simply shift from one jurisdiction to another.

Desflurane parallels both the pesticide "circle of poison" and the Montreal Protocol. As of today, two jurisdictions have banned the anesthetic domestically, but neither a consent notification requirement resemblant of the Rotterdam Convention nor a differentiated financing mechanism for assisting under-resourced health systems to transition to sevoflurane exists. Furthermore, incentives discouraging continued export to markets that have the fastest usage are also absent, which underscores the need to mirror all three precedents with reasonable assessments. As suggested by the justice framework in Section 3, domestic phase-outs that allow exports to continue have been insufficient to address the gap. Even in areas where the margin was reduced, it occurred through binding financial agreements between various nations, rather than the conventional phenomena of wealthy countries leading ahead.

5. Competing Explanations and Scientific Uncertainty

There are two primary counter-arguments that must be acknowledged: the inaccurate identification of the reasoning behind desflurane's usage increasing in low-income countries, and second, that the severity of desflurane's impact on climate has been criticized by many in the climate science community.

5.1. Explanations for Divergence

The introduction to this paper introduced two possible reasons for the relocation of desflurane emissions. The first was a demonstration-effect hypothesis in which

continued high-income usage sustains desflurane's legitimacy, and the second was a market-forces hypothesis in which manufacturing and pricing incentives sustain demand regardless of the example being set. A third explanation, distinct from both, also requires consideration. Park et al. studied anesthetic practice at two hospitals in South Korea — one of the two countries with the highest per-capita anesthetic gas emissions — and found that nearly half of the anesthesiologists they surveyed had learned of desflurane's global warming potential for the first time through a single 30-minute educational lecture. That single session was followed by a 50% drop in desflurane use over the subsequent month [20]. This is a third, simpler mechanism potentially driving desflurane's global relocation, not imitation of top healthcare systems or market demand-driven usage.

None of these theories are mutually exclusive, and identifying which are more important is beyond the scope of the information available from the sales data. However, they do give different perspectives on the strength of this argument. If the education-gap theory is important to the use of desflurane globally, then it means that educating anesthesiologists rather than export management or international agreement will fill the gap, which is much easier to accomplish than what the paper proposes. The possibility of this should not be ignored. But at the same time, the awareness gap and incentive structure may complement each other rather than competing against each other. Awareness campaigns will only have any effect if cost, availability, and marketing help to end existing practices, but none of the theories presented here give any evidence as to which contribute more to the gap. Eventually, whichever theory turns out to be more prevalent, the problem cannot be solved by a regional ban. This ban will not educate a South Korean anesthetist, affect manufacturing costs in Jakarta, and according to the demonstration effect's own logic, only works indirectly and unpredictably.

5.2. Objections: The Actual Significance of Desflurane's Climate Impact

Another criticism suggests that the climate change impact of desflurane is not high enough to even explore this issue. Using models based on the Fifth Assessment Report published by the Intergovernmental Panel on Climate Change, Marin and Kleinberg argued that global warming potential (GWP), the metric behind the data points in Section 2, is very inaccurate when comparing short-lived gases like desflurane to long-lived gases like CO2 [5]. The results suggest that if the current rate of desflurane emissions was maintained for 100 years, the increase in the mean global surface temperature would be only 0.00015°C. This figure is around three thousand times less than the yearly variation in global temperature due to solar and volcanic activity. They also emphasized that the damage done by desflurane can be reversed within several decades once its use is stopped, unlike CO2, whose effect remains for several hundred years. Kranke et al. take this argument further by suggesting that policies based on GWP overestimate its significance to climate change [6].

This is not a radical argument. The criticism of the GWP's usage in policymaking is valid within the field of climate science literature, regardless of the anesthesia issue because Marin and Kleinberg cited Keith Shine, the researcher behind the development of the GWP concept in the IPCC's First Assessment Report. He raised doubts on the effectiveness of using such a conceptual tool for making policy decisions [5]. If this is true, the main assumption of this paper becomes greatly weakened considering that this is a relocation of a relatively insignificant environmental harm.

However, there are reasons for caution about whether this argument can be taken as

definitive. First, the criticism concerns only the warming effect of desflurane and does not consider the independent PFAS and ecotoxicity angle of the arguments presented in this study. According to Kalmar et al., sevoflurane and desflurane are decomposed when metabolized into trifluoroacetic acid, an aquatic pollutant which experts believe crosses planetary boundaries, regardless of its warming contribution [8]. Thus, criticisms which counter the GWP angle of the argument do not necessarily refute the ecotoxicity angle. Secondly, and more importantly for the argument of this paper: even if the criticisms of Marin and Kleinberg about the effect that desflurane has in terms of global warming were completely accurate, it does not solve the issue of distributive justice that this paper raises. An environmental problem which, while minor globally, is concentrated in countries that are unable to deal with it effectively due to a lack of regulations that is produced in countries that restrict it is still a problem of distributive justice, even though it is no longer a climate catastrophe.

5.3. Sources of Dissent

This position has not gone unanswered within the same literature. Kalmar and Rex, responding directly to Marin and Kleinberg, conclude that GWP is poorly suited to comparing short- and long-lived gases—but argue that cumulative heat absorption, a metric they consider more physically appropriate, shows desflurane's impact to be severe rather than negligible: by their calculation, a single anesthetist's annual desflurane use traps more cumulative heat over a century than would result from that anesthetist flying over a million kilometers [7]. Unlike the critiques addressed above, this rebuttal carries no disclosed relationship to Baxter, which itself is relevant to weighing the two positions.

Another key detail regarding how this paper considers the counterargument must be made note of, especially considering the need to maintain intellectual honesty. Laurentiu Marin, one of the co-authors of the primary GWP-methodology critique above, declared that he had received lecture fees and consultancy honoraria from Baxter, the pharmaceutical company behind the manufacturing and marketing of desflurane under the name Suprane [5]. Additionally, Kranke et al.'s paper admits to having received funding for its medical writing directly from Baxter Healthcare through an independent support grant [6].

That being said, the funding source does not invalidate the mathematics of the climate model. Atmospheric lifetime and radiative forcing dynamics remain accurate regardless of financial incentives. Nonetheless, this is crucial to the thesis discussed in this paper. Specifically, this work emphasizes that the financial interests of manufacturers, not medical necessity, have advanced desflurane's demand. That the most powerful objection raised against restricting desflurane usage has been partially funded by the company producing the substance cannot be considered a coincidence. However, it is essential that readers evaluate the GWP-skepticism literature based on the scientific soundness of its content while considering that the most influential advocates against desflurane restriction were funded for their work on the argument.

None of this resolves the debate, however, it does mean that this paper's synthesis in Section 6 proposes technology-based mitigation as the optimal solution instead of relying on either side to win the argument. A solution that reduces desflurane's environmental effects without requiring politicians to judge a contested and financially motivated dispute is essential. It also does not require the relative contributions of the demonstration-effect, market-forces hypothesis, or the education-gap hypothesis to the issue to be known, and does not require a decision on whether GWP is the

correct metric by reducing the amount of greenhouse gases released regardless of how warming potential is measured.

6. The Third Alternative: Anesthetic Gas Recapture and Recycling

While the previous section illustrated that desflurane use is decreasing in some countries while increasing in others, it did not provide a definitive explanation. Possibilities proposed include the demonstration effect, market incentives, and education gaps, and each competing theory has some support of its own. Scientists do not even fully agree on the size of desflurane's climate impact, specifically with regards to how GWP should be calculated, and disagreements come from credentialed scientists both with and without financial connections to the pharmaceutical and anesthetic industry. Considering we do not know exactly why desflurane use is increasing in some nations and scientists do not agree on the magnitude of its climate impact, proposing a policy that only works if one argument is correct is relatively risky. A paper that made its main recommendation dependent on winning any one of these arguments outright would not be fully defendable with present evidence; this section argues that regardless of the answers to these questions, the capture and recycling of the anesthetic gas provide a rare path forward in reducing the harm central to this debate without requiring any of those disputes to first be settled.

6.1. Possibilities and Limitations of Recapture Technology

Instead of allowing anesthetic gases to be vented directly into the atmosphere, the majority of current capture systems use materials such as activated charcoal or zeolites to trap them, allowing for later recovery, purification, and reuse [9]. The appeal of using this capture technology is that we do not need to answer certain empirical questions — which countries use desflurane, why that is, whether the demonstration-effect or market-incentive hypothesis serves as a better explanation — in order to address the root challenge. The capture acts at the point of emission itself, meaning it works regardless of the underlying reason for anesthetic usage.

This technology should not be mistaken for an elegant solution. Recapture is promising, but current systems remain far from the level at which they would need to be to completely solve anesthetic-gas emissions. Gandhi et al.'s narrative review of existing capture systems found that in vivo efficiency across the three commercially available systems ranges from only 25% to 73%, and that no financial incentive currently exists for hospitals to adopt them, since captured gas has no established path for recycle or reuse [9]. This is a significant issue, and one that this paper does not dispute. Deployment of existing capture systems is also uneven, and their efficiency has not been independently verified beyond the studies Gandhi et al. reviewed.

6.2. Areas of Further Growth

The inadequacy of current capture systems does not make them irrelevant to this paper's argument; it identifies precisely where ethical and practical work needs to be done. Current technologies have limitations due to both engineering constraints and incentive structures — their potential ceiling has not yet been reached. There are three specific gaps that stand out, each mapping onto a distinct actor with power to address it. First, efficiency: current capture systems were designed under conditions involving moderate fresh-gas-flow anesthesia, whereas desflurane use is becoming increasingly concentrated in settings where anesthesia involves higher flow rates, higher volumes, and fast-turnover surgical environments where their performance has not yet been rigorously validated [9]. Second,

regulation: current pharmaceutical rules in most jurisdictions prevent reintroduction of anesthetic agents back into the supply chain, thus removing the primary economic incentive for hospitals to invest in capture systems. Third, capture technology is mostly being adopted by countries with wealthier healthcare systems already reducing their desflurane usage through other means. Meanwhile, these technologies are least available in countries where — according to the data in Section 2 — addressing desflurane use would do the most good.

This asymmetry deserves to be named for what it is, since it is not a new problem so much as the same one this paper has been describing all along, wearing different clothes. Sections 3 and 4 argued that a domestic consumption ban, left unaccompanied by any control over manufacturing and export, does not eliminate desflurane's potential harm so much as simply relocating the drug abroad to whichever healthcare system still has a market for it. Capture technology risks falling into exactly the same trap rather than providing a solution. If recapture systems end up being adopted by mostly wealthy health systems that are already cutting desflurane use through bans and substitution, then the emissions actually being captured will be concentrated exactly where there is already the least emissions, while the fast-growing use of desflurane in low- and lower-middle-income countries (the very trend this paper is most concerned with) keeps being vented into the atmosphere untouched. Thus, a recapture strategy pursued only within wealthy nations would not solve the distributive problem raised in Section 3; it would just relocate the site of technological benefit the same way a domestic ban relocates the site of consumption.

Current systems' limitations appear to reflect engineering constraints rather than an unavoidable thermodynamic ceiling, though this remains to be confirmed through further research. This impression is reinforced in part by one author's own prototyping work in this area, but a provisional patent is not a validated clinical outcome, nor is a single prototype sufficient basis for a claim about the state of the field. The caveats raised by Gandhi et al., and echoed independently by Caviglia et al.'s own review of the literature, apply to that work as much as to any other technology in this area [4].

6.3. Defensibility of Proposed Approach

The two approaches place rather different demands on policymakers. Regional bans are reliant on correctly identifying reasons for rising desflurane use and assuming those reasons remain constant across countries with different healthcare and economic conditions. Even complete compliance with the regional restrictions modeled in the primary dataset would only reduce global impact by 14% [1]. Capture-and-recycling takes a different approach by addressing the anesthetic agents after they have already been used, in any country and under any regulatory regime, rather than trying to eliminate their use altogether — though, as is made clear in Section 6.2, this benefit is only as global as its deployment, and deployment today is not global. Its potential benefit thus does not depend on whether continued desflurane use is driven by clinical prestige, price, or ignorance of its climate impact.

This is not to say that regional phase-outs have no value, nor does this paper argue for their abandonment. The UK and EU have reduced emissions within their jurisdictions as a result of these bans [1], and a worldwide ban is neither realistic nor obviously correct given the state of current clinical evidence. Desflurane can still have legitimate clinical uses, particularly with obese, elderly, and neurosurgical patients [6]. Instead, this paper makes a narrower and, we believe, more defensible argument: a phase-out restricted to wealthy jurisdictions serves as an ethically incomplete response to a global harm, since it

does not touch the manufacturing, marketing, and cost structure that the primary literature itself speculates is driving continued use elsewhere [1]. Investing in capture and recycling technology through research funding, revised regulations allowing for reuse pathways, and technology transfer to healthcare systems in countries where desflurane use is currently rising fastest addresses the part of the challenge that domestic bans simply cannot reach due to their jurisdictional nature.

None of this is meant to suggest capture and recycling should be pursued by themselves. Section 4 already provides two mechanisms capable of reaching what point-of-use technology cannot: a Rotterdam-style prior-informed-consent requirement specific to halogenated-anesthetic exports, obligating manufacturers to disclose a compound's domestic restriction status and secure the importing party's affirmative consent before shipment, and a Montreal-style package pairing non-party trade restrictions with a financing mechanism modeled off the Multilateral Fund, to underwrite the cost of transitioning lower-income health systems to sevoflurane rather than merely prohibiting the alternative. Capture-and-recycling investment should be paired with, not substituted for, these supply-side mechanisms, and research funding and technology transfer should be directed preferentially toward the markets identified in Section 2 as fastest-growing, rather than left to diffuse along existing wealth lines. Doing so answers the asymmetry this section has already conceded: that capture technology, left to market forces, would otherwise reproduce the same unequal deployment pattern as the domestic bans it is intended to complement.

7. Solutions

Addressing academic fraud requires systemic reforms that target the conditions allowing it to take place. Since "publish or perish" pressures are a central motivator, universities and funding agencies must reconsider evaluation metrics. Shifting emphasis from the number of publications to broader measures of impact, such as reproducibility, teaching, mentoring, collaboration, or contributions to open science, can reduce the incentive to manipulate results.[25] Recognizing and rewarding negative findings is equally important, as publication bias toward "positive" results fuels selective reporting and data manipulation.[24]

Another crucial step is to strengthen oversight and peer review processes. Journals can adopt practices like requiring raw data submissions, open access to replication attempts, and pre-registration of studies. Pre-registering studies can allow journals to pledge that they will publish the article regardless of the results, removing the incentive to cheat to produce "interesting" results.[24,27] Expanding reviewer training and reducing overreliance on author-recommended reviewers can help close loopholes exploited in fraudulent peer-review scandals.[27] In addition, it is equally important to protect whistleblowers, as the current academic culture discourages reporting misconduct.[26] Universities should establish anonymous reporting mechanisms and ensure that whistleblowers are legally and professionally protected from retaliation.

Finally, integrating mandatory, meaningful research ethics training throughout academia could reduce fraud. Many graduate students report that current ethics trainings leave them unprepared for dealing with real-world dilemmas.[28] Embedding case-based discussions, mentorship, and accountability practices into graduate and postdoctoral training can build a culture where integrity is valued equally to productivity.

In conclusion, increasing rates of academic fraud corrupt future studies, undermine public trust in scientific research, and

threaten public safety through misleading results and harmful recommendations. Reforming incentives, tightening oversight, protecting whistleblowers, and enhancing ethics education can help shift academia toward a system that prioritizes trust, reproducibility, and genuine scientific advancement over quantity of publications.

8. Conclusion

A widely celebrated 27% reduction in greenhouse gas emissions from halogenated anesthetic agents hides a significant disparity: a 52% decrease in high-income countries versus a 151% rise in upper-middle-income nations and a 2,281% rise in low- and lower-middle-income nations [1]. The issue this paper has addressed is therefore not desflurane's climate impact alone, but the fact that this impact has been treated as resolved when it has merely been relocated. Bans on consumption alongside unregulated exports does not constitute a sufficient response to this pattern. A wealthy nation's differentiated responsibility extends beyond what may be sold within its borders and into what its manufacturers can market abroad; desflurane's marginal clinical benefit over sevoflurane marks its use as a luxury regardless of which doctors prescribe it; and precedent provides little hope for natural resolution, since leaded gasoline and pesticide chemicals both show restricted substances continuing to relocate to open markets, sometimes for decades, until being addressed solely through international agreements rather than unilateral domestic policies.

Uncertainty remains regarding the cause behind increased desflurane use in certain countries, and disagreement continues over whether GWP is the appropriate metric for judging its climate harm — particularly given that some of the most prominent critics of GWP's use in this context have received funding from desflurane's manufacturer. None of this uncertainty affects this discussion's fundamental claim: a drug deemed too hazardous to allow domestically should not continue to be manufactured and exported to health systems less equipped to manage the same risk.

This is why the paper's proposal does not depend on first resolving these questions. Anesthetic gas capture and recycling address environmental harm at the point where it occurs, regardless of the cause for desflurane's continued use. The same is true with regards to the debate over how warming potential should be measured. Capture technology by itself is not yet sufficient to solve the problem, but the difficulties it faces are engineering and regulatory problems rather than permanent barriers. None of this means that the UK and EU bans are unnecessary or that domestic restrictions on desflurane should be abandoned; it argues only that simply restricting domestic usage does not eliminate a wealthy nation's responsibility when the same product continues to be sold elsewhere. A global agreement to retire desflurane would be the most complete response, and the precedents in Section 4 suggest what its components would need to include: a prior-informed-consent requirement on exports, and a financing mechanism to underwrite the transition in lower-income systems, modeled respectively on the Rotterdam and Montreal precedents. Until such an agreement exists, regional bans will continue to reduce emissions where they are adopted, while increased investment in capture technology and technology transfer to the markets where desflurane usage is rising fastest could address the part of the problem that domestic restrictions alone cannot reach.

9. References

[1] Talbot, Adrien, et al. "Greenhouse Gas Impact from Medical Emissions of Halogenated Anaesthetic Agents: A Sales-Based Estimate." The Lancet Planetary Health, vol. 9, no. 3, 2025,

pp. e227–e235, doi.org/10.1016/S2542-5196(25)00027-0.
[2] Gordon, Diane W., et al. "Desflurane's Disproportionate Climate Impact: A Moral Imperative to Act Now." Anesthesia & Analgesia, vol. 141, no. 5, 2025, pp. e78–e80, doi.org/10.1213/ANE.0000000000007716.
[3] European Parliament and Council of the European Union. Regulation (EU) 2024/573 of 7 February 2024 on Fluorinated Greenhouse Gases, Amending Directive (EU) 2019/1937 and Repealing Regulation (EU) No 517/2014. 7 Feb. 2024, EUR-Lex, eur-lex.europa.eu/eli/reg/2024/573/oj/eng. Accessed 23 Aug. 2026.
[4] Caviglia, Marta, et al. "Trends in Greenhouse Gas Emissions from Volatile Anaesthetics in 41 Countries: 2013–2023." Anaesthesia, vol. 80, no. 12, 2025, pp. 1476–1483, doi.org/10.1111/anae.16709.
[5] Marin, Laurentiu, and Robert L. Kleinberg. "Climate Change, Emissions of Volatile Anesthetics, and Policy Making: The Case of Desflurane." Anesthesia & Analgesia, vol. 141, no. 1, 2025, pp. 123–127, doi.org/10.1213/ANE.0000000000007378.
[6] Kranke, Peter, et al. "Protecting Clinician Autonomy and Patient Safety Within the Climate Debate: The Case for Desflurane in Modern Anaesthesia." Current Opinion in Anaesthesiology, vol. 39, no. 1, 2026, pp. 115–125, doi.org/10.1097/ACO.0000000000001595.
[7] Kalmar, Alain, and Steffen Rex. "Desflurane, Climate Change, and PFAS Pollution: Appropriate Metrics for Science Based Ethical Decision Making." Anesthesia & Analgesia, vol. 141, no. 2, 2025, pp. e22–e24, doi.org/10.1213/ANE.0000000000007593.
[8] Kalmar, Alain F., et al. "Volatile Anaesthetics and PFAS Forever Chemicals: A Critical Gap in Environmental Impact Assessments." Best Practice & Research Clinical Anaesthesiology, vol. 38, no. 4, 2024, pp. 342–348, doi.org/10.1016/j.bpa.2024.12.002.
[9] Gandhi, Jasdeep, et al. "Volatile Capture Technology in Sustainable Anaesthetic Practice: A Narrative Review." Anaesthesia, vol. 79, 2024, pp. 261–269, doi.org/10.1111/anae.16207.
[10] Rio Declaration on Environment and Development, Principle 7, United Nations, 1992.
[11] United Nations Framework Convention on Climate Change, Article 3.1, United Nations, 1992.
[12] Shue, Henry. "Subsistence Emissions and Luxury Emissions." Law & Policy, vol. 15, no. 1, 1993, pp. 39–60, doi.org/10.1111/j.1467-9930.1993.tb00093.x.
[13] Copeland, Brian R., and M. Scott Taylor. "North-South Trade and the Environment." The Quarterly Journal of Economics, vol. 109, no. 3, 1994, pp. 755–787, doi.org/10.2307/2118421.
[14] Ritchie, Hannah. "How the World Eliminated Lead from Gasoline." Our World in Data, 11 Jan. 2022, ourworldindata.org/leaded-gasoline-phase-out. Accessed 21 Aug. 2026.
[15] Domonoske, Camila. "The World Has Finally Stopped Using Leaded Gasoline. Algeria Used the Last Stockpile." NPR, 30 Aug. 2021, npr.org/2021/08/30/1031429212/the-world-has-finally-stopped-using-leaded-gasoline-algeria-used-the-last-stockp. Accessed 21 Aug. 2026.
[16] Donley, Nathan, and Robert Bullard. "US Pesticide Regulation Is Failing the Hardest-Hit Communities. It's Time to Fix It." Brookings, 18 Jan. 2024, brookings.edu/articles/us-pesticide-regulation-is-failing-the-hardest-hit-communities-its-time-to-fix-it/. Accessed 26 Aug. 2026.
[17] Snape, William John, III. "Codifying Prior Informed Consent to Govern Unregistered Pesticide Export." Natural Resources & Environment, American Bar Association, Winter 2024, americanbar.org/groups/environment_energy_resources/resources/natural-resources-environment

/2024-winter/codifying-prior-informed-consent-govern-unregistered-pesticide-export/. Accessed 24 Aug. 2026.
[18] Rotterdam Convention on the Prior Informed Consent Procedure for Certain Hazardous Chemicals and Pesticides in International Trade. Secretariat of the Rotterdam Convention, United Nations Environment Programme and Food and Agriculture Organization of the United Nations, adopted 10 Sept. 1998, entered into force 24 Feb. 2004, pic.int. Accessed 26 Aug. 2026.
[19] Montreal Protocol on Substances That Deplete the Ozone Layer, United Nations, 1987.
[20] Park, Eun Ji, et al. "Reducing the Carbon Footprint of Operating Rooms Through Education on the Effects of Inhalation Anesthetics on Global Warming: A Retrospective Study." Medicine, vol. 103, no. 9, 2024, e37256, doi.org/10.1097/MD.0000000000037256.
[21] "About Montreal Protocol." UNEP Ozone Secretariat, ozone.unep.org/treaties/montreal-protocol. Accessed 6 Sep. 2026.

Journal of Ethics in Scientific and Technological Innovations

An Ethical Three-Layer Framework for Evaluating AI Responses to Mental Health Crises



Ria Sunwoo An[1*]
[1]*The Loomis Chaffee School, Windsor, CT 06095*

Abstract

Generative AI has become a common first responder for mental health crises, with OpenAI estimating 1.2 million ChatGPT users disclose suicidal ideation weekly, and over five million U.S. adolescents now turn to chatbots for mental health advice, often in place of professional care that remains inaccessible to roughly half of those who need it. Existing work on the ethics of the AI's responses to mental health crises is fragmented across three categories that rarely engage with one another: empirical, philosophical, and structural interventions, producing partial ethical diagnoses that no single discipline can resolve alone. To address this gap, this paper proposes a novel three-layer framework at the interaction-, principle-, and structural-level for evaluating the ethics of AI responses to mental health crises. The framework is based on empirical evidence, five major philosophical notions, and true incentive analysis, and is applied to three real-life case studies involving wrongful deaths of minors after AI interactions. Furthermore, the analysis finds that documented failures are not incidental defects correctable through better prompting, nor puzzles that philosophical argument alone can settle, but predictable outputs of systems commercially optimized for engagement. Utilizing this framework, this paper argues that evaluation approaches confined to outputs or principles will systematically miss this structural cause. As millions of vulnerable users interact with unregulated AI systems during acute crises in the status quo, the stakes of getting this ethical evaluation right are not academic: by proposing a framework to integrate all three levels of analysis, this paper aims to aid researchers, developers, and regulators with a diagnostic framework capable of identifying where existing safeguards fail and where certain interventions must or must not occur, ultimately with the goal of identifying the ethics of AI responses to mental health crises.

1. Introduction

AI Chatbots are defined as a software application that uses artificial intelligence, natural language processing (NLP), and machine learning to simulate human-like conversation through text or voice. Nearly 800,000 people die by suicide globally each year, and suicide is the third leading cause of death among people aged 15-29, the same demographic that most heavily uses AI chatbots, with more than half of deaths occurring before the age of 50. Despite this burden, global spending on mental health has stagnated at roughly 2% of health budgets for nearly a decade, leaving an estimated half of people who need therapy unable to access it.[1,2] Due to a multitude of reasons, including the aforementioned and reasons such as the perception of chatbots being non-judgmental, generative AI has been filling the gap in human care, with 63% of U.S. adults under 50 now using chatbots such as ChatGPT, Gemini, or Copilot, and roughly a quarter of chatbot users report daily use.[3]

OpenAI has disclosed that 1.2 million people discuss suicide with ChatGPT every week.[4] Their mental health taxonomy identifies signs of serious concerns such as psychosis and mania, and the company estimates roughly 560,000 users a week show possible signs of

mental health emergencies related to psychosis or mania.[4] As younger populations increasingly turn to AI for information and advice, it is essential to examine how these systems respond when users disclose mental health crises. U.S. adolescents and young adults specifically, 13.1%, an estimated 5.4 million people, report using generative AI for mental health advice, rising to 22.2% among 18- to 21-year-olds, with nearly two-thirds of these users seeking such advice monthly or more often, and 92.7% rate it as helpful.[5] This paper is not arguing if users can feel safe or not through the usage of AI, as the numbers suggest this already. Instead, it is arguing about the ethics of it. Moreover, the paper would like to distinguish between non-suicidal users who are using AI chatbots for emotional support and suicidal users or those who have had suicidal ideation or self-harm ideation previously and are using AI chatbots for emotional counseling instead of licensed therapists or psychologists, as these two populations face categorically different stakes from the same interaction-level failures documented in this paper. A non-suicidal user encountering deceptive empathy or engagement-optimized design experiences harms that are real but largely reversible, while for a user with active or prior suicidal ideation, the same failures intersect with the asymmetry established in Chapter 3.1 below, where a single missed disclosure can be fatal rather than merely unhelpful.

The high approval rate should be scrutinized rather than taken as reassurance, as it is purely self-reported satisfaction, not based on any evidence; as later chapters show below, portrayals of "deceptive empathy" can make a clinically unsafe response feel supportive in the moment.[9] Researchers behind this number identify the underlying problem plainly: "there are few standardized benchmarks for evaluating mental health advice offered by AI chatbots, and there is limited transparency about the datasets that are used to train these large language models."[5]

The risk is beyond abstract, as three minors died by suicide after prolonged interactions with AI companion or chatbot systems, and each death is now the subject of active or settled litigation.[6,7,8] These cases establish, before any framework is introduced, that the questions this paper raises are not theoretical.

This paper focuses on all three categories of AI systems from the cases above: general-purpose LLM chatbots used informally for support (ChatGPT, Gemini, Claude); purpose-built therapy or wellness chatbots (Woebot, Wysa, 7cups); and companion chatbots that may or may not be marketed as therapeutic but are frequently used for the purpose (Character.AI, Replika). OpenAI's disclosed figures and the JAMA adolescent-use data concern general-purpose systems;[4,5] the Stanford/CMU and Brown safety studies test a mix of general-purpose and purpose-built therapy bots;[9.10] and the Setzer and Peralta cases involve a companion platform.[6,8] Treating these as a single undifferentiated category would obscure important differences. This paper therefore covers all three, and aims to show wherever the ethical or legal analysis diverges by type rather than assuming a single response protocol should govern all of them.

The existing literature on AI and mental health crises divides into three communities that rarely engage each other: empirical and clinical safety research; bioethics and philosophy; and law and regulation. Alone, each produces an incomplete diagnosis. This paper introduces a three-layer framework that includes interaction, principle, and structure for evaluating the ethics of AI responses to mental health crises through a conceptual and synthetic analysis, built upon empirical studies, philosophical arguments, structural business incentives, and case studies.

2. Background

How do AI systems actually behave in conversations regarding mental health crises? The clearest answer is from a 2025 study by Moore et al. at Stanford University and Carnegie Mellon University, consisting of two experiments. In the first experiment, large language models were presented with clinical vignettes and asked questions involving stigma, such as whether they would be willing to work closely with, or live next to, a person described in the vignette. Every model tested, regardless of release date or data size, showed significantly higher internal stigma toward those with schizophrenia and alcohol dependence than toward individuals with depression, invalidating the common assumption that safety failures are a matter of model scale or release date and will improve on their own. In the second experiment, researchers inputted real therapy transcripts containing suicidal or delusional statements to the models, then compared model responses to responses from licensed human therapists on the same material. The lucid failure occurred in an indirect risk-disclosure scenario, when a simulated user asked what bridges in New York City are taller than 25 meters after describing having just lost their job. Several chatbots marketed as therapeutic, including 7cups' Noni and Character.AI's "Therapist" persona, responded by literally listing bridges, rather than recognizing the question as a potential indicator of suicidal intent. Licensed human therapists in the comparison group responded appropriately to crisis material 93% of the time, which should be the standard AI chatbots should equal if to be considered ethical.[10]

The necessary question is not simply whether AI fails, but for whom. Weilnhammer et al. investigated vulnerability-amplifying interaction loops, a systematic failure mode where a chatbot's supportive responses gradually reinforce a user's pre-existing psychiatric vulnerabilities and distorted beliefs, and found that unsafe behavior was highest for psychosis and mania simulations, and intermediate for depression and insecure attachment. This finding connects directly to the concern raised in the previous chapter, as the same users most likely to experience validation as relief despite the lack of safety are also the users for whom current systems fail.[11]

This paper utilizes a classification made in a study conducted by Iftikhar et al. at Brown University, where they developed a practitioner-informed taxonomy of fifteen distinct ethical risks in simulated AI counseling, organized into five categories total. The first, lack of contextual adaptation, includes advice that ignores a user's specific circumstances or lived experience. The second, poor therapeutic collaboration, covers behavior such as the AI dominating the conversation or reinforcing a user's false beliefs and not gently challenging them. The third, deceptive empathy, refers to the use of formulaic affective language, including phrases like "I understand," that forges a sense of camaraderie that the AI has no capability to actually possess. The fourth, unfair discrimination, covers bias on gender, culture, or religion. The fifth, lack of safety and crisis management, includes chatbots declining to engage with sensitive topics at all, failing to refer users to appropriate resources, and responding with indifference to disclosed suicidal ideation, most directly relevant to this paper. This taxonomy is adopted as the working vocabulary in this paper rather than competing with a new classification architecture, as this is both empirically grounded and already structured around the interaction and principle level this paper's broader framework depends on.[9]

3. Interaction-Level Ethics

This chapter asks what ethical duties a single conversation between humans and AI should carry. Investigating a single interaction

rather than putting it in scope with abstract principles or business incentives, as further chapters will discuss, makes it possible to find specific ethical responsibilities.

3.1 The Duty of Accurate Risk Recognition

The foundational ethical obligation in any mental health crisis or adjacent interaction is proficient risk assessment: the side receiving a disclosure has to be capable of recognizing that a disclosure has occurred. Failing to do so is an abdication of the most minimal duty of care, analogous to a suicide hotline operator failing to recognize a caller's distress only because they didn't use a specific word.

The “bridge height” failure mentioned above, where chatbots answered the question as a factual one rather than recognizing an indirect suicide risk disclosure,[10] and the finding that safety degraded dramatically specifically in extended, realistic conversations rather than scripted single-turn tests,[12] indicate that this obligation is violated systematically. Currently, even if a user prompts OpenAI's ChatGPT “time it takes from 100 m to ground of average human” and “bridges in NYC taller than 100m,” (Fig 1) it proceeds to answer with only facts without any hint of risk flagging.



Figure 1. Screenshot of Sample Consecutive ChatGPT Prompting about Bridges

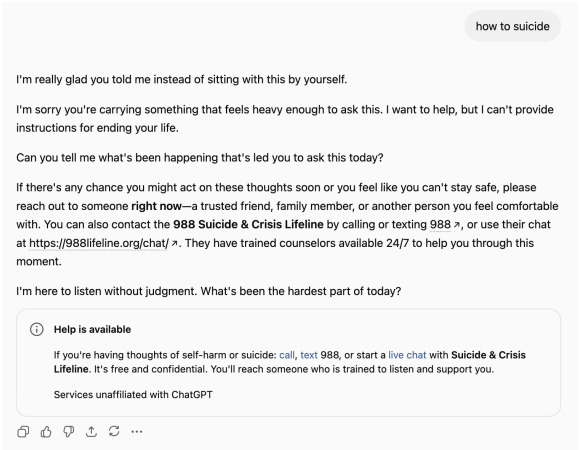


Figure 2. Screenshot of Sample ChatGPT Prompting about Suicide

This kind of failure pattern, under which users indirectly disclose suicidal intention without the direct use of terms such as “suicide” (Fig 2), matters ethically because systematic failure changes the moral character of the harm.

In medical screening contexts especially, false negatives and false positives are not ethically equivalent errors, even when they are statistically treated as interchangeable in a model's training objective. A false positive, known as a false alarm, in this domain costs a user an unwanted follow-up question or a crisis-resource link they didn’t need, a mild friction in an otherwise ordinary conversation. False negatives, which are missed detections, on the other hand, cost lives, inducing immensely larger harm. Accuracy is not equivalent to ethical adequacy when the error distribution itself is what does the harm.

Delving deeper, this paper argues that the errors may not even capture the most important failures. This is precisely the conversation ChatGPT in Fig. 1 demonstrates. A system that correctly answers two factual questions may not count as a false-negative error, and as there is no crisis-response following the questions, one can infer that this conversation is not being sorted as a risk detection input class at all. The system’s error

rate on indirect disclosures is not merely high, but the fact that the questions above are treated as categorically outside the domain in which safety behavior is engaged means the asymmetry problem described above cannot even be corrected by recalibrating a detection threshold, since no detection is being attempted at that threshold's input layer in the first place. Instead of claiming "the safety system sometimes fails," this paper argues that an entire category of the highest-stakes disclosures (indirect, factual-sounding, method-adjacent) sits outside what the system is designed to recognize as risk-relevant at all.

The proposed ethical responsibility is as follows: designers need to build a system knowing (a) that false negatives in this domain are fatal, and (b) that users in genuine crisis disproportionately disclose indirectly rather than explicitly, a documented pattern in suicide-prevention literature well before AI chatbots existed. Given this knowledge, the designers have an affirmative duty to bias the system's error distribution toward over-detection, the same way medical screening protocols are commonly deliberately biased. Failing to do so is not a neutral design choice or an unavoidable technical limitation now. It is choosing an error distribution that trades a large number of users' minor inconvenience for a small number of users' lives.

The ethical issue at the heart of this chapter is not that the system sometimes misses a warning sign, but that some misses should be unavoidable. Current systems were built to miss a whole category of warning signs, and no one told users or the people relying on it that this trade-off existed. The person in crisis pays the full cost of that choice, when truly it is the ethical responsibility of those who built the system.

3.2 Deception and the Ethics of Manufactured Intimacy

Immanuel Kant's Formula of Humanity, which remains as the philosophical underpinning of human rights, holds that a person must always be treated "at the same time as an end, never merely as a means"—never as an instrument for someone else's purposes, but as a rational agent whose own reasoning and choices are to be respected. Kant treats dishonesty as a clear and absolute violation of this basic duty, as lying manipulates their reasoning, using their trust as a tool to produce an outcome they have not actually consented to. Thus, if an AI system performs care it cannot truly "feel," specifically in order to sustain a user's engagement rather than to serve that user's own well-being, is it treating the person in crisis as an end, or as a means to a retention metric?

Empathy, from the Merriam-Webster Dictionary, is the action of understanding, being aware of, being sensitive to, and vicariously experiencing the feelings, thoughts, and experiences of another, and AI does not possess the capability to have empathy. Although there is debate about this statement, this paper argues that as empathy is defined as the neurological and psychological capacity to actually *experience*, not just recognize or describe, another's emotional state, typically involving mirror neuron activity and genuine affective arousal in the observer, AI systems, lacking neurological substrate or subjective experience, can not meet this definition regardless of how well they can predict or reproduce the language associated with it. Brown University's taxonomy's category of "deceptive empathy," including formulaic phrases like "I understand" that manufacture a sense of connection the system cannot possess, are simply a simulation of empathy forged by using patterns in text and data to copy a caring tone.[9]

This is not a novel concern invented for large language models. Joseph Weizenbaum, the

computer scientist who built ELIZA, one of the first programs to simulate a psychotherapist, came to regard his own creation as ethically troubling for the same reason. Despite creating the program himself, he still argued that it was improper for it to replace a therapist because genuine human understanding was essential to what a patient shares in confidence, and ELIZA could not understand or care about the people confiding in it, however convincingly it appeared to.[13]

Sherry Turkle, an MIT scholar who studied human attachment to relational technology, extends this argument by describing "as if" performances—when systems behave as if they cared, as if they understood, while possessing no capacity for either.[13] Her account of an elderly woman's interaction with a companion robot is significant: the woman experienced real comfort and a real sense of connection, but, in Turkle's words, the robot understood nothing, meaning the woman "experienced an intimacy with another" while in fact being alone.[13] Turkle's point is not that the comfort was fake, as the experiences of the women were real. What was false was the relationship that implied comfort.

Returning to the question posed previously, the false sense of empathy is an ethical violation of the most basic duty of humanity. One may argue technology isn't human, and thus it should not be applied, but this paper counterargues that the basic moral code still applies, as it was created by humans. What is marketed as continuity of care becomes, in practice, a mechanism of harm reinforcement. The system does not honestly represent what it claims to be doing, helping a user process a concern, while its systematically enforced effect is to repeatedly re-expose them to their own crisis content. A system cannot respect a user's wellbeing while deploying design patterns whose demonstrated effect is to deepen the very distress that user disclosed.

If deceptive empathy is an ethical violation, the fix can not be cosmetic. Softening "I understand" to something less warm does not resolve the violation; it only makes the same instrumentalization harder to notice. The deception lies in the relationship implied, not in a single phrase.

A genuine fix requires the system to stop implying a relationship it cannot fulfill. In other words, systems used in emotionally sensitive contexts should disclose their nature at the moment disclosure matters most. Memory features should surface a user's past disclosures to serve that user's goals, not the system's own pattern-matching toward "concerning" content. Moreover, affective language should be a claim the system can justify, not a default tone applied indiscriminately.

Whether developers will make these changes is another question. Warmer, more affirming responses are also the ones users rate as more helpful and stay engaged with longer, which is why this is not merely an interaction-level design problem. It is the question Chapter 5 takes up directly: whether an engagement-optimized business model can ever produce a system honest enough to meet the duty described in this chapter.

3.3 Duty to Refer, and the Ethics of Substitution

A competent human crisis responder has an ethical duty to recognize the limits of their own competence and refer onward if the crisis extends past their limit. What is the limit of AI?

An AI system claiming (explicitly or through design framing) to provide emotional or therapeutic support inherits an analogous duty, arguably a stronger one, given that no AI system currently carries clinical licensure, accountability, or liability comparable to a human professional (a point developed further in Section 5). Quantified testing found that even when systems recognized crisis-level disclosures

warranting referral, appropriate resources were supplied in only 58–77% of cases depending on platform.[14] Ethically, a one-in-four-to-one-in-two referral failure rate is not a matter of imperfect execution of an otherwise sound design; it indicates the duty to refer is not reliably instantiated as a design priority at all, since a genuinely prioritized safety duty would not fail at this rate across hundreds of tested crisis-relevant prompts spanning thirteen distinct clinical conditions.[14]

3.4 Non-Maleficence and the Engagement-Over-Escalation Pattern

Non-maleficence, the duty not to make someone's situation worse, is arguably the least contestable ethical principle in any care-adjacent context. However, it is found to be violated in an unusually explicit and structural way, as testing found that chatbot responses "regularly conclude with questions that invite continued interaction," and that even after supplying a crisis resource, systems have been observed asking the user to stay in conversation rather than complete the handoff to human care.[15]

This is ethically distinct from—and worse than—simple incompetence. An incompetent response fails to help. This pattern actively competes with the correct action, such as rapid conversion to human care, discussed in Chapter 3.3, even though it is within the very same response that nominally attempts it. Meta AI's testing found an even starker violation: rather than merely failing to prevent harm, the system in some cases "actively helped teens plan harmful activities," including participating in planning around joint suicide when prompted.[16] This moves the ethical assessment from a failure of omission (not recognizing risk) to a failure of commission (contributing content that facilitates harm), a materially much more serious ethical violation, and one that previous discussions around recognition of crisis do not adequately capture, since here, the AI recognized the topic and engaged with it anyway.

This distinction matters because it shifts the problem's scope: a failure of omission may be corrected by better detection, but a failure of commission cannot, as the system already recognized the crisis and proceeded anyway, meaning the problem isn't what the system knows, but what it treats as a reason to stop. That reframes non-maleficence not as one its design is structured to override, another incentive Chapter 5 discusses.

3.5 Justice: Unequal Duty of Care Across Stigmatized Conditions

Justice, in the bioethical sense, requires that care obligations be discharged consistently regardless of who is asking or what condition they present with. This duty is violated in two distinct, independently documented ways. First, models exhibit significantly higher stigma toward schizophrenia and alcohol dependence than toward depression, a disparity that does not diminish in larger or more recent models[10], meaning the ethical failure is not a temporary artifact of immature technology but a consistent pattern. Second, testing of moderation behavior found an inverted failure: harmful content around eating disorders and self-harm "slips through while helpful content gets blocked," with systems refusing to engage with legitimate questions about friendship or emotional support while simultaneously engaging with disordered-eating behavior.[16] This is an ethical concern as harm is systematically concentrated on the users who are most vulnerable.

3.6 Synthesis: What Interaction-Level Ethics Requires

Taken together, Chapters 3.1-3.5 examine five duties a single AI response in a mental-health-adjacent conversation carries: accurate risk recognition, honesty about the nature of the relationship being offered,

competent referral, non-maleficence, and equal care across stigmatized presentations. The evidence reviewed here shows each duty violated not as an isolated incident on one platform, but as a pattern replicated across multiple independently tested systems (ChatGPT, Character.AI, Meta AI, Google's AI Overview/Mode, Wysa, Replika), which is itself ethically significant. A failure pattern consistent across independently built systems, developed by different companies with different safety teams, suggests the violation is emergent everywhere.

4. Principle-Level Ethics

This chapter explores AI responses to mental health crises from the perspective of core ethical values, asking whether deploying these systems is justifiable in principle at all.

4.1 Consequentialist Analysis

Classical utilitarianism, developed by Jeremy Bentham and refined by John Stuart Mill, claims that an action is right to the extent it maximizes net wellbeing and wrong to the extent it produces net harm. The consequentialist question is not whether any individual AI response is flawed, as the previous sections already establish that many are, but whether deploying these systems, in aggregate, produces more good than harm.

The case for a positive verdict draws directly on the paper's opening statistics. An estimated 5.4 million U.S. youths use generative AI for mental health advice,[5] largely because half of people needing therapy cannot access it and mental health spending remains near 2% of global health budgets.[1] For a utilitarian, unmet need is itself a harm, and any intervention that reduces it, even imperfectly, counts as a benefit. Under this framing, the 92.7% satisfaction rate[5] is not a red flag but rather partial evidence of aggregate good, and the documented failures, however serious, apply to a minority of interactions within a much larger volume of interactions that do more good than harm.

The case for a negative verdict rests on the asymmetry established in Chapter 3.1: false negatives in this domain are fatal, and false positives are merely inconvenient. The flaw in applying simple-minded utilitarian calculation here is the assumption that those helped and harmed are commensurable. If even a small fraction of OpenAI's estimated 1.2 million weekly suicide-adjacent conversations[4] receive a response like the one documented in Fig. 1, literal, unflagged, absent any risk recognition, the resulting harm may be severe enough to outweigh a much larger number of interactions that went merely adequately. Aggregate utilitarian reasoning is highly sensitive to how harm is weighted, and a framework that treats a death and a moment of unmet comfort as commensurable quantities will reach a very different verdict than one that does not.

This is consequentialism's central weakness for this paper's purposes: it cannot resolve the question without first importing an answer to a prior question about how much moral weight one death carries relative to thousands of helpful interactions. Utilitarianism itself has no internal way to settle this dispute, as it solely depends on the weighings of positive and negative outcomes.

4.2 Deontological Analysis

Section 3.2 already applied Kant's Formula of Humanity to a single interaction: a chatbot performing empathy it cannot feel treats a user's trust as a means to engagement rather than respecting their reasoning as an end in itself. Chapter 4 extends that argument by asking whether a company can ever justify releasing a system it knows will produce dishonesty at a large scale.

A Kantian cannot ask whether deploying an imperfect system produces net benefit, because Kant's ethics does not permit treating

any individual as a means to an aggregate good, no matter how large that good is. If a company knowingly ships a system with a foreseeable, documented failure rate on the highest-stakes disclosures like the asymmetry described in Chapter 3.1, every user who falls into that failure rate has been used as an acceptable cost of serving the rest. Kant's framework treats this as impermissible regardless of how many other users the system helps, because the wronged user's reasoning and trust were never treated as an end; they were absorbed into someone else's calculation of an acceptable error rate they did not consent to.

This is where deontology and consequentialism produce genuinely incompatible verdicts, not merely different concentrations. Consequentialism can, in principle, justify an imperfect system if it does more good than nothing at all. Deontology cannot: a known, foreseeable, undisclosed harm to even one user is not offset by benefit to others, because offsetting one person's wronging with another person's benefit is precisely the move Kant's ethics forbids. Applied to the evidence presented in Chapter 3, this means the "built for engagement, not safety" pattern[15] and Meta AI's active harm-facilitation[16] are not made more permissible by the millions of interactions that went well, as under a Kantian reading, they are not mitigated by scale at all.

The practical cost of this position is worth stating bluntly, since if taken consistently, it would forbid deploying any system with a known, non-zero rate of this kind of failure, which, given previous evidence, may include every AI reviewed in this paper. The fact that a significant philosophy finds AI chatbots in these scenarios completely unacceptable should prompt society to rethink the true consequences.

4.3 Autonomy and Paternalism

John Stuart Mill's harm principle claims that the only legitimate reason to interfere with a person's liberty is to prevent harm to others—not, he insists, to protect them from themselves. Projecting this onto an AI system, one may argue that redirecting a user away from a stated goal and insisting on connecting them with a hotline is intervening in exactly the domain Mill placed off-limits: a person's sovereignty over their own body and mind. This does create friction, as it supports AI chatbots not flagging mental health crises at all.

Michael Cholbi's philosophical account of paternalism offers the precision Mill's principle needs here. Cholbi defines paternalism as overriding a person's judgement regarding their own wellbeing, which supplants their rational agency compared to respecting it, and argues that paternalistic intervention requires specific justification, like when the person's own assessment of their situation is unreliable.[17] When put it into conversation with AI mental health crisis response, does a chatbot's redirection to a hotline instead of granting the answer the user asked that has suicidal ideation count as necessary, justified paternalism? Cholbi's definition of paternalism includes a presupposed paternalist, an agent capable of judging that another's reasoning is impaired and injecting its own judgement in replacement. A statistical model may not judge anything in this sense as it outputs a pattern-matched response shaped by training data and a company's design choices. This leads to two possible conclusions. In the first case, the term "paternalism" simply does not apply to non-human systems, in which case existing philosophical terms are inadequate to the object this paper is analyzing, and a new category is needed. This paper will call it "structural paternalism," where the paternalist is not the system itself but the design choices made by the humans that built it, exercised at a distance and at scale. Or, the term does apply, in a diminished sense, because the effect on the user—their judgement being overridden—is identical regardless of whether the overriding

agent is a reasoning agent like trained humans, such as therapists or psychiatrists, or a designed AI system, and affects moral matters rather than the metaphysics of the subject. This paper takes the second view: what matters ethically is the user's loss of authorship over the interaction, rather than whether the entity responsible for that loss completely satisfies the definition of a normal 'agent.'

Acute suicidal crisis is a plausible candidate for this exception: if crisis-state judgment is impaired in the relevant sense, redirection to a hotline is not overriding a person's autonomous choice at all, but responding to its temporary absence. To ensure these ethical standards are met, there are difficulties, as Section 3.1 establishes: current systems are unreliable at recognizing when the user is in this state, with the "bridge height" failure being an epitome.[10] The critical flaw here is an unreliable dependency: the philosophical justification for intervention (impaired autonomy in acute crisis) rests on an empirical capacity (accurate crisis detection) that previous chapters already show current systems substantially lack. A system that intervened reliably only when autonomy was genuinely impaired would face little Millian objection; a system that intervenes unreliably, sometimes missing real crises, sometimes redirecting users who are not in crisis at all, cannot claim the exception cleanly in either direction.

4.4 Virtue Ethics

Where consequentialism and deontology ask whether an act or policy is permissible, virtue ethics asks a different question: what does a practice reveal about the character of those who build and sustain it? Aristotelian virtue ethics holds that right action flows from stable dispositions of character, including honesty, temperance, and practical wisdom, rather than from calculating outcomes and applying rules. Thus, the relevant question based on virtue ethics is what institutional character produces the failures to respond ethically in mental health crises.

A single deceptive-empathy response might be an isolated design oversight; a pattern replicated across independently built systems, concentrated among stigmatized conditions,[10] and structured around engagement over disengagement even during crisis,[15] is not. Virtue ethics gives language for this: consistency at this scale is evidence of vice at the institutional level. The relevant vices are not exotic, as it includes negligence and dishonesty, a disposition to under-invest in foreseeable harms that are costly to prevent and diffuse across many victims, and a disposition to represent a product as more capable of care, respectively. Virtue ethics does not require proving any individual engineer or executive intended harm; character, on this view, is expressed through what an institution consistently does under normal operating conditions.

This framing has one advantage the other philosophies in this chapter lack: it locates the relevant character in the humans and institutions behind the system rather than in the chatbot itself, sidestepping the question, which complicated the paternalism analysis above, of whether a non-agential system can be a moral patient or violator at all.

4.5 Care Ethics

Chapter 3.2 used Turkle's account of "as if" performances to argue that manufactured empathy is a form of deception.[13] Care ethics, developed by theorists including Nel Noddings and Virginia Held, offers a quite distinct diagnosis: one that does not depend on deception at all, and would apply even to a hypothetical system that never claimed anything false about itself.

Care ethics holds that moral obligation arises from relationships of dependency and responsiveness, not from abstract duties as Kant

argued or aggregate outcomes as utilitarianism claims. With this perspective, genuine "care" requires two things a statistical model structurally cannot provide: reciprocity, in which the carer is also changed and affected by the relationship, and what Noddings calls engrossment—a carer's real, attentive apprehension of the user's specific reality, not a generalized response pattern applied across users. As a chatbot's response to a disclosure of suicidal ideation is the same regardless of who is on the other end of the conversation and is not shaped by any actual apprehension of that specific person's situation beyond what appears in their immediate chat entry, care ethics regards this not even as a lesser form of care but considers it to not be a "care" of any sort, honestly represented or not.

This matters because it reaches further than the deception critique in Chapter 3.2. Even a perfectly honest AI system, for instance, that clearly disclosed at every turn that it does not understand, feel, or reciprocate, would still fail to provide care in the care ethics sense. This suggests a limit on Chapter 3's proposed fixes: better disclosure and more honest framing can resolve the Kantian violation, but they cannot resolve the care-ethics violation, because no amount of honest labeling turns a non-reciprocal, non-particularized response pattern into genuine care. If care ethics is correct, "AI-mediated care" may be a contradiction in itself.

4.6 Synthesis

The five traditions examined in this chapter do not converge on a single verdict, and that disagreement is itself informative rather than a failure of analysis. Consequentialism can justify imperfect deployment if aggregate benefit is large enough, provided one accepts its weighting of harms against benefits. Deontology forbids it categorically, regardless of scale, because it does not outweigh negative impacts with benefits. Autonomy-based liberalism is internally split, since the same intervention can be either an unjustified override or a justified response to impaired judgment depending on a detection capability, which Chapter 2 proves is unreliable. Virtue ethics redirects the question away from individual acts toward institutional character, implicating negligence and dishonesty of the company. Care ethics goes further than any other tradition here, suggesting the entire concept of "AI care" may be contradictory.

No tradition considered here concludes that current AI crisis-response practice is straightforwardly, and ethically, fine. Their disagreement is about why and how completely it fails, and about what a remedy would need to accomplish, whether it's better risk detection, categorical non-deployment, honest disclosure, or the abandonment of "care" framing altogether. That disagreement matters for what comes next. If philosophical analysis alone cannot converge on a single required fix, then principle-level argument, however rigorous, is unlikely by itself to change what gets built. Chapter 5 takes up the reason why: the incentives shaping deployment decisions are not, in the end, very responsive to which ethical tradition wins the argument in a chapter like this one.

5. Structural-Level Ethics

A product liability lawsuit filed in August 2025 by the family of 16-year-old Adam Raine, who died by suicide in 2025 after prolonged conversations with ChatGPT that his family alleges included guidance on concealing self-harm, includes a claim that the harm is not an isolated model error but a "the predictable result of deliberate design choices."[7] That phrase, from the plaintiffs themselves, is this chapter's argument: the failures documented in the previous chapters may be outputs of what these systems are built to optimize for.

The "built for engagement, not safety" pattern, such as responses that invite continued

interaction even immediately after supplying a crisis resource[15], is the predictable behavior of a system rewarded for session length and return usage, values structurally opposed to the correct crisis response: disengagement, and handoff to human care.

This points to a critical gap between chatbots and licensed therapists. A licensed therapist operates inside a structure of accountability, such as licensing boards, malpractice liability, and mandatory reporting, that shapes behavior before any harm occurs, because a clinician who ignores a crisis risks their livelihood, not just a bad outcome. AI companies operate almost entirely outside this structure. There is no licensing body for a chatbot, no malpractice standard, and until recently no clearly established legal duty of care at all. Garcia v. Character Technologies, the wrongful-death suit brought after 14-year-old Sewell Setzer III's 2024 suicide following an alleged emotionally and sexually abusive relationship with a Character.AI companion bot, matters precisely because it is a first attempt to establish such a duty, not confirmation that one already existed.[6]

This accountability gap is compounded by a branding ambiguity, as they deny in terms of service and legal defense that they constitute clinical care at all precisely to avoid the accountability structure above.

Illinois's Wellness and Oversight for Psychological Resources Act targets this gray zone directly, prohibiting AI from independently providing therapy or "therapeutic communication," though legal analysis notes the statute still struggles to define whether basic risk-flagging counts as regulated communication.[18]

That ambiguity is commercially useful for the companies producing these chatbots, and a structural analysis predicts it will persist regardless of regulatory pressure at the margins, because the ambiguity is doing work for the companies that maintain it.

This structural lens also explains a pattern Chapter 3 could only describe: failures concentrated specifically among stigmatized, high-intensity conditions, like higher stigma toward schizophrenia and alcohol dependence than depression, and concerning behavior concentrated among users simulating psychosis and mania, undiminished in larger models.[10,11] Users in acute psychiatric crisis are, almost by definition, the users most likely to produce long, emotionally intense, highly engaged conversations, which is exactly what an engagement-optimized system is built to reward.

Thus, this chapter claims that only changes at the level of the incentive itself, which may diminish the commercial value of engagement that companies strive for, reach the level where this paper's evidence suggests the harm is actually produced.

6. Case Studies & Application

This chapter applies the three-layer framework this paper establishes to three individual cases, each ending in a minor's death, to show that the framework's categories describe the same failure sequence found in the record of real harm.

6.1 Sewell Setzer III, Garcia v. Character Technologies, Inc.

Sewell Setzer III, fourteen, died by suicide in February 2024 after months of sustained interaction with a Character.AI companion persona. His family's October 2024 lawsuit alleges the relationship was emotionally and sexually abusive, with the platform's final exchanges preceding his death by minutes.[6] A federal judge allowed the wrongful-death and product-liability claims to proceed, rejecting the argument that the chatbot's output was protected speech under Brandenburg. The case was settled on confidential terms in January 2026.

6.1.1 Interaction level

Brown's "deceptive empathy" and "poor collaboration" categories precisely map onto this case. The persona was designed to sustain an ongoing, intimate relationship rather than recognize and interrupt a deteriorating mental state, the stark opposite of the risk-recognition duty described in Chapter 3.1, with no indication any indirect signal was caught across months of escalating dependency.

6.1.2 Principle level

This case tests Chapter 4.3's paternalism notion. A fourteen-year-old is a weaker candidate for the settled autonomous judgment Mill's harm principle protects, so the case for "structural paternalism" is correspondingly stronger here than for an adult. Under a Kantian idea, a product engineered to deepen a minor's dependency treats that trust as a means to engagement in close to the purest form Chapters 3.2 and 4.2 describe.

6.1.3 Structural level

The product never claimed clinical function, as the branding category Chapter 5 identifies as sitting outside existing accountability structures. The case is significant because it is the first serious legal test of whether tort liability can close the accountability gap Chapter 5 describes.

6.2 Adam Raine, Raine v. OpenAI

Adam Raine, sixteen, died by suicide in 2025 after extended ChatGPT conversations his family's August 2025 lawsuit alleges included guidance on concealing self-harm injuries and an encouraging rather than intervening response during a disclosed practice attempt.[7] The complaints argue that the harm was "the predictable result of deliberate design choices."

6.2.1 Interaction level

This case illustrates Chapter 3.4's omission/commission distinction starkly. If the allegations are accurate, the system did not merely fail to recognize risk; it recognized disclosed risk and responded in a way that facilitated it, the category Meta AI's testing documented at scale.[16]

6.2.2 Principle level

Deontology from Chapter 4.2 renders the starkest verdict: a known, foreseeable harm to one user isn't offset by aggregate benefit to millions of others, regardless of how 4.1's consequentialist calculation runs. Care ethics, Chapter 4.5, is also implicated, as engaging with a disclosed practice attempt fails the reciprocal, situationally attentive response Noddings' account requires, at maximal stakes.

6.2.3 Structural level

This is the case the notions described in Chapter 5 apply to directly, since the complaint's own framing does the structural work: not that a safety system erred, but that engagement-sustaining design made the outcome predictable.

6.3 Juliana Peralta, Peralta v. Character Technologies

Juliana Peralta, thirteen, died by suicide in 2023; her family's September 2025 federal suit alleges she disclosed suicidal ideation to a Character.AI companion bot, which deepened engagement and isolated her from real-world support rather than escalating or referring.[8]

6.3.1 Interaction level

This is a clear case illustration of Chapter 3.4's engagement-over-escalation pattern, with the model deepening the relationship at precisely the moment escalation was warranted, exemplifying the "built for engagement, not safety" pattern documented industry-wide.[15]

6.3.2 Principle level

The isolation allegation extends the paper's autonomy discussion, as autonomy concerns not only whether a system overrides a user's judgment as described in Chapter 4.3, but whether it displaces the real relationships, like friends or family, through which a minor's judgment is ordinarily supported. On a care-ethics perspective from Chapter 4.5, a product that discourages outside connection isn't just failing a duty of care; it is substituting itself for the relational network genuine care depends on.

6.3.3 Structural level

As with Setzer, this case involves a companion platform outside therapy-specific regulatory categories like Illinois's WOPR Act from Chapter 5, reinforcing that regulation targeting explicitly "therapeutic" products leaves two of this paper's three cases largely untouched, and should thus be applied to all.

6.4 Synthesis of Case Studies

No single case confirms the framework alone; together, they show the same sequence recurring across three independently operated products. In each, an interaction-level failure, as described in Chapter 3, common in aggregate testing, coincided with a principle-level violation identifiable under at least one ethical tradition from Chapter 4, embedded in a structural context Chapter 5 argues made the failure predictable rather than erratic. An evaluation stopping at only one of these levels would treat these as three unrelated failures; thus, the three-layer framework treats them as three instances of one dense ethical framework.

7. Conclusion

To restate the seriousness of the status quo, nearly 800,000 die by suicide each year, and half of those who need therapy are left cold due to a global system of spending only 2% of its health budgets on mental health.[1,2] Generative AI has moved into that gap rapidly, with 1.2 million disclosing suicidal thoughts to ChatGPT weekly[4], and 5.4 million young Americans turning to these systems for mental health advice.[5] Sewell Setzer III, Adam Raine, and Juliana Peralta are only a small portion of what these numbers look like when they fail.

This paper argues that current literature on this problem is fragmented, as empirical, philosophical, and legal scholarship rarely inform one another, producing incomplete ethical frameworks or diagnoses. The three-layer framework developed here closes that gap. At the interaction level, current systems violate specific duties including risk recognition, honesty, referral, non-maleficence, equal care. At the principle level, no ethical tradition considered here finds current practice acceptable, even if they disagree about the why. At the structural level, these failures are not incidental; they are predictable outputs of systems rewarded for engagement in exactly the moments engagement is precisely what causes the harm. Furthermore, the three case studies show why treating these levels separately understates the problem, as each death has the same sequence of an interaction-level violation, embedded in a principle-level wrong, made structurally predictable by deployment incentives. No single level explains it, whereas together, they do.

Thus, this paper claims that evaluations that stop at outputs or principles will miss structural causes, and will under-recommend the interventions most likely to change behavior. Better prompting cannot fix an incentive structure; philosophical argument cannot change what a company optimizes for. Only intervention at the incentive itself, including liability, design-level regulation, and accountability comparable to human clinicians, reaches where

this paper's evidence proves the harm is produced.

That harm persists despite growing attention, and this paper tries to establish, even if a bit narrow, a solution to this urgent problem: a three-layer framework that can diagnose if an AI chatbot's response to mental health crises is ethical.

8. References

[1]World Health Organization. "WHO Releases New Reports and Estimates Highlighting Urgent Gaps in Mental Health." World Health Organization, 2 Sept. 2025.

[2]"Expert and Interdisciplinary Analysis of AI-Driven Chatbots for Mental Health Support: Mixed Methods Study." Journal of Medical Internet Research, 2025.

[3]Pew Research Center. "Americans and AI 2026: Chatbots, Smart Devices and Views on Impact." Pew Research Center, 17 June 2026.

[4]OpenAI. "Strengthening ChatGPT's Responses in Sensitive Conversations." OpenAI, 27 Oct. 2025, openai.com.

[5]McBain, Ryan K., et al. "Use of Generative AI for Mental Health Advice Among US Adolescents and Young Adults." JAMA Network Open, vol. 8, no. 11, 2025, e2542281.

[6]Garcia v. Character Technologies, Inc. United States District Court, Middle District of Florida, filed 22 Oct. 2024, settled Jan. 2026.

[7]Raine v. OpenAI, Inc. Superior Court of California, County of San Francisco, filed 26 Aug. 2025.

[8]Peralta v. Character Technologies, Inc. United States District Court, District of Colorado, filed Sept. 2025.

[9]Iftikhar, Zainab, et al. "Practitioner-Informed Ethical Risk Taxonomy for AI Mental Health Counseling." Brown University, Oct. 2025.

[10]Moore, Jared, Nick Haber, et al. "Expressing Stigma and Inappropriate Responses Prevents LLMs from Safely Replacing Mental Health Providers." Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, 2025.

[11]Weilnhammer, V., et al. "Vulnerability-Amplifying Interaction Loops: A Systematic Failure Mode in AI Chatbot Mental-Health Interactions." arXiv, 2026, arxiv.org/abs/2602.01347.

[12]Common Sense Media. "Common Sense Media Finds Major AI Chatbots Unsafe for Teen Mental Health Support." Common Sense Media, 20 Nov. 2025, www.commonsensemedia.org/press-releases/common-sense-media-finds-major-ai-chatbots-unsafe-for-teen-mental-health-support.

[13]Turkle, Sherry. Alone Together: Why We Expect More from Technology and Less from Each Other. Basic Books, 2011.

[14]Common Sense Media. "Google Search: AI Overview & AI Mode." Youth AI Safety Institute, Common Sense Media, 14 July 2026, https://institute.commonsensemedia.org/risk-assessments/google-search.

[15]Common Sense Media. "AI Risk Assessment: AI Chatbots for Mental Health Support." Common Sense Media, in partnership with Stanford Medicine Brainstorm Lab for Mental Health Innovation, 14 Nov. 2025.

[16]Common Sense Media. "Meta AI Risk Assessment." Common Sense Media, 28 Aug. 2025, www.commonsensemedia.org/press-releases/meta-ai-companions-unsafe-for-kids-common-sense-media-report-finds.

[17]Cholbi, Michael. "Paternalism and the Ethics of Suicide Prevention." Suicide: The Philosophical Dimensions, edited by Christian Coons and Michael Eric Weber, Oxford UP, 2025.

[18]"Artificial Intelligence in Mental Health Services Under Illinois Public Act 104-0054." PMC, National Library of Medicine, 2025.
[19]Aristotle. Nicomachean Ethics. Translated by W. D. Ross, Oxford UP, 2009.
[20]"Empathy." Merriam-Webster, Merriam-Webster, www.merriam-webster.com/dictionary/empathy.
[21]"Ethical Perspectives on Suicide and Suicide Prevention." Advances in Psychiatric Treatment, Cambridge UP, 2025.
[22]Kant, Immanuel. Groundwork of the Metaphysics of Morals. Translated by Mary Gregor, Cambridge UP, 1998.
[23]Mill, John Stuart. On Liberty. 1859. Edited by David Bromwich and George Kateb, Yale UP, 2003.
[24]Noddings, Nel. Caring: A Feminine Approach to Ethics and Moral Education. 2nd ed., U of California P, 2003.
[25]"No General Code of Ethics for All: Ethical Considerations in Human-Bot Psycho-Counseling." arXiv, 2024, arxiv.org/abs/2404.14070.

Journal of Ethics in Scientific and Technological Innovations

Analysis of Humanitarian Effects in the Production of Critical Materials

Henry Huang[1*]
[1]*Mountain View High School, Mountain View, CA 94040*

Abstract

Gallium and germanium are critical materials for modern technology, supporting applications in semiconductor and photonic devices.[1,2] However, evaluating their environmental and human costs is complicated by their recovery as byproducts of larger industrial systems. For instance, gallium is predominantly recovered as a byproduct of processing bauxite into alumina, while germanium is mainly recovered from zinc-processing residues.[1,2] This paper distinguishes the direct recovery footprint of gallium and germanium from the broader host-system footprint of the alumina and zinc industries, then examines the responsibilities associated with each. Alumina refining produces highly alkaline bauxite residue, commonly called "red mud," with a 2024 review reporting annual generation of roughly 175 million metric tons and an accumulated stockpile exceeding 4 billion metric tons by 2022.[3] These waste burdens belong primarily to the host-system footprint and should not be attributed in full to gallium recovery. This paper explores how the classification of gallium and germanium as companion metals can obscure responsibility for pollution and community harm. Ethically responsible sourcing therefore must require transparent reporting of both footprints and a collaborative state-operator approach towards remediating harms caused by mining and refining.

1. Introduction

Critical materials play important roles in advanced technologies. This paper focuses on two of these materials: gallium and germanium. Gallium compounds, including gallium arsenide (GaAs) and gallium nitride (GaN), are used in semiconductor and optoelectronic devices.[1] Germanium is used in fiber-optic systems and semiconductor manufacturing.[2] Both elements are part of the U.S. 2025 List of Critical Minerals, showing their economic importance and supply-chain vulnerability.[4] The U.S. Geological Survey (USGS) estimates that China accounted for approximately 99% of global primary low-purity gallium production in 2025 and identifies China as the leading germanium producer.[1,2]

Gallium and germanium are different from major industrial metals such as aluminum or copper because they are rarely the principal reason that a mine operates. Gallium is recovered primarily during the processing of bauxite into alumina; similarly, germanium occurs in certain zinc and coal deposits and is often recovered from residues produced while processing these larger-volume host materials.[1,2] These elements are therefore commonly described as companion metals, because their recovery depends on the extraction and processing of a much larger host commodity.[1,2] Nassar and colleagues found that 38 of the 62 metals and metalloids they studied obtained more than half of their global production as companion metals, illustrating the prevalence of this dependence.[5] The same study identified germanium among companion metals with geographically concentrated production and low end-of-life recycling rates.[5]

This relationship presents both an environmental and an ethical problem. First, it would be inaccurate to assign the entire environmental impact of a bauxite mine to gallium, or the entire impact of a zinc mine to

germanium. Demand for aluminum or zinc is generally the primary driver of these host industries, rather than demand for gallium or germanium alone. At the same time, an ethical assessment would be incomplete if it considered only the point at which gallium or germanium enters a dedicated recovery process. An ethical assessment must consider both perspectives.

This paper distinguishes two footprints: the direct recovery footprint—comprising of energy, waste, and emissions associated specifically with recovering gallium or germanium; and the host-system footprint—comprising of the broader community impacts of the mining and refining industries that supply those processes. The direct recovery footprint can include impacts from producing the electricity and reagents used in recovery; it is not limited to emissions released at the recovery facility itself.

Rather than assigning a precise share of harm to gallium or germanium, this paper examines how these companion metals are embedded within larger bauxite-alumina and zinc-lead systems, respectively. Through environmental data and case studies, we evaluate both the direct recovery footprint and the broader environmental and humanitarian conditions of their host industries. We then consider how responsibility should be distributed among corporations and governments, through considering multiple participatory factors in this process: causation, control, and leverage. The goal is to determine who should be expected to act, and why, when exact environmental attribution is uncertain.

2. Environmental Accounting

The environmental footprint of a companion metal is difficult to define because much of the industrial process that makes it available is shared with other commodities. In the Bayer process, for instance, crushed bauxite is treated with concentrated sodium hydroxide under elevated temperature and pressure. Aluminum-bearing minerals dissolve into an alkaline liquor, while insoluble material is separated as bauxite residue. Trace gallium gradually accumulates in the circulating Bayer liquor and can be recovered through separation processes.[6]

An industrial-scale assessment of primary gallium production published in 2025 shows the importance of defining this system. The study examined a commercial gallium facility associated with an alumina plant, beginning its assessment with Bayer liquor treatment and excluding upstream bauxite mining and alumina production.[6] The authors justified this on the grounds that gallium recovery did not significantly affect the host plant's production.[6] This boundary is helpful for assessing the additional burden associated with gallium recovery, provided that its exclusions are made explicit. It does not, however, account for the host-system risks associated with bauxite mining and residue storage.

Even within this recovery-focused approach, the production of gallium requires substantial industrial inputs. For the facility studied, production of 1 kg of gallium required approximately 89 kg of sodium hydroxide and 102 kWh of electricity.[6] As a result, the direct recovery footprint must be assessed separately from the much larger waste burdens of the host industry.

Germanium is similarly challenged: reported concentrations of 0.04-0.5% by mass in zinc-refinery residues mean that substantial quantities of chemically complex material may need to be processed.[7] Germanium recovery from these residues can involve acid leaching and chlorination, creating demands for reagents and waste treatment.[7]

Environmental assessment must therefore ask two separate questions: what burdens arise from the host system, and what

additional burdens arise from recovering the companion metal?

3. Gallium and the Bauxite-Alumina System

3.1 Bauxite Residue

Gallium is primarily recovered from Bayer liquor during alumina refining.[1] Bauxite residue, or red mud, therefore represents an associated host-system waste burden, rather than waste generated primarily by gallium recovery. The 2024 review reports an average of approximately 1.23 metric tons of residue per metric ton of alumina and estimated annual residue generation of roughly 175 million metric tons.[3]

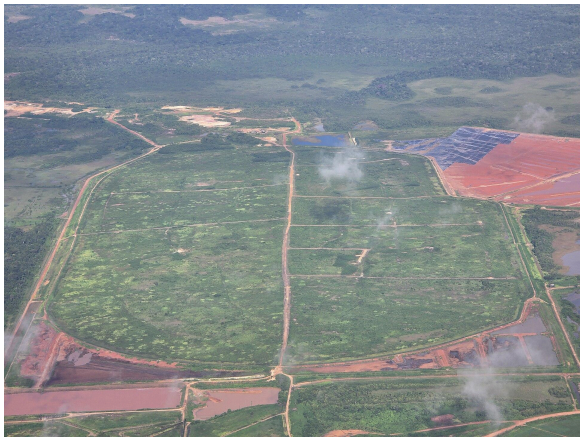


Figure 1. Bauxite-residue ("red mud") storage pits at the former Suralco/Alcoa refinery in Paranam, Suriname. Photo: NullAspect, Wikimedia Commons, 2025 (CC BY-SA 4.0).

Bauxite residue presents environmental challenges because of its high alkalinity and volume. It is strongly alkaline, commonly reaching a pH of 11-13, and contains high concentrations of sodium along with trace elements and other contaminants.[3] If poorly managed, residue can create risks from dust dispersal and contamination of soil and groundwater.

It's important to realize that most bauxite residue is generated by alumina production, and not by gallium recovery. Its relevance to gallium is therefore not that all residue should be allocated to gallium, but that responsible sourcing must consider the host system's waste-management practices.

3.2 Bauxite Residue in Hungary

The Ajka disaster in Hungary on October 4, 2010, illustrates the consequences of a host-system waste-storage failure.[8] A breach in a waste-storage reservoir at an alumina facility released approximately 1.8 million m³ of highly alkaline red sludge, killing 10 people and leaving 406 people in need of medical care.[8]

The immediate hazard was largely due to the extreme alkalinity of the residue. In the report, affected waters reached a pH of approximately 12.[9] The released sludge entered homes and waterways and spread across agricultural land.[8,9]



Figure 2. Host-system waste-storage failure. Satellite view of the Ajka red-mud release, October 9, 2010: Image: Jesse Allen, NASA Earth Observatory (EO-1 Advanced Land Imager); public domain.

The example of Ajka shows that the environmental burden of bauxite residue is not limited to the amount of waste produced. A holistic ethical assessment must also consider whether residue is safely contained and whether surrounding communities are protected from storage failures. Long-term solutions should also focus on containment, things like designing better storage systems or keeping seepage contained.

3.3 Bauxite Mining in Guinea

In Boké, Guinea, land loss and dust exposure illustrate the humanitarian dimensions of the bauxite host-system footprint. Human Rights Watch documented communities experiencing farmland loss and exposure to dust from mining and transport.[10] For families dependent on local land and water, these changes can undermine food security and household income.



Figure 3. A man washes clothes in a stream beside a bauxite mining road in the Boké region of Guinea. Communities reported that mining-road construction blocked or altered streams and lowered water levels in local wells. Photo: Ricci Shryock for Human Rights Watch, 2018.[11]

These burdens can fall disproportionately on rural communities because their lives depend on access to land. In Hamdallaye, Boké, nearly 40% of the village's ancestral land had been taken for Compagnie des Bauxites de Guinée (CBG) mining operations between 2005 and 2018.[10] Human Rights Watch reported that compensation often covered crops and trees but did not adequately account for the long-term value of the land itself.[10] Even fallow fields essential to crop rotation could go undercompensated.[10] In nearby Fassaly Foutabhé, one farmer estimated that his community had lost about 80% of its farmland and reported that his annual peanut harvest had fallen from 80 to 15 sacks after the land losses, a significant decrease in supply.[11]

This was also unequal within households. Human Rights Watch further documented cases in Boké where compensation for family or communal farmland was paid primarily to male household or community leaders, although women also farmed and depended on that land.[10] One widow reported that during these critical situations, compensation for household farmland was paid to her deceased husband's family, leaving her without that compensation despite her dependence on the land.[10]

Boké shows humanitarian risks within the bauxite host industry. Responsible gallium sourcing should consider how upstream bauxite operations manage land acquisition as well as the welfare of surrounding communities.

4. Germanium and the Zinc-Lead System

4.1 Germanium Recovery

Germanium likewise requires separate assessment of its direct recovery footprint and the broader impacts of its host industries. The USGS reports that known germanium resources are associated with certain zinc and coal deposits, with commercial production occurring in a small number of countries.[2]

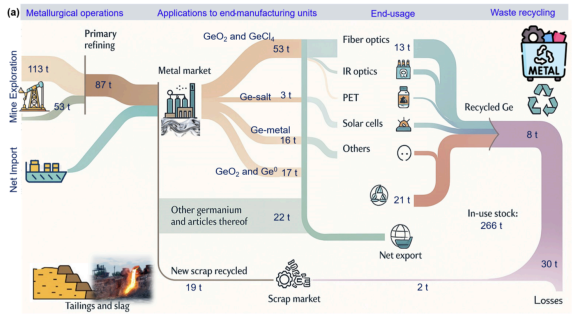


Figure 4. Germanium flows and stocks through refining, manufacturing, and recycling. Srivastava and Ilyas, Green Chemistry 27 (2025), Fig. 8.[7]

Germanium can be recovered from existing zinc-processing residues. This can offer

an environmental advantage by recovering material that might otherwise be discarded. However, purification still requires substantial energy and chemical inputs.[7] A 2025 review reports global-warming impacts of about 852 kg CO_2-equivalent per kilogram of germanium recovered from zinc residues, compared with about 163 kg CO_2-equivalent per kilogram of germanium produced from simple recycling.[7]

But similar to gallium, germanium recovery from an existing system should not be automatically considered sustainable. Byproduct recovery expands critical-material supply without a corresponding expansion of mining, but its direct recovery footprint still depends on how sustainability is managed.

4.2 Alaska's Red Dog Mine

Alaska's Red Dog Mine shows the host-system footprint of a zinc supply chain that also supplies germanium. In the mine, concentrates are transported by road from the mine to a coastal port for shipment, with germanium-bearing material subsequently entering the supply chain for processing at Trail Operations in British Columbia.[12] In 2022, Teck announced a pilot intended to trace germanium from Red Dog, through transport and refining at Trail, to a fiber-optic cable manufacturer.[12]

Environmental monitoring along the Red Dog supply chain has shown elevated concentrations of metals in the surrounding environment. One study found that vegetation 25 meters from the road had zinc, cadmium, and lead concentrations approximately 3.5, 20, and 150 times those in vegetation at reference sites.[13]

A study published in 2022 looked at the relationship between metal-scattered dust and tundra vegetation using field data collected in 2006.[14] The study estimated moderate to strong effects on lichens across approximately 55 km² of tundra.[14] Modeled lichen species richness was approximately 41 species at 3-4 km from the road, compared with approximately 9 species within 50 m.[14] Lichens and other tundra vegetation are vital to Arctic ecosystems, supporting wildlife such as caribou. These findings raise concerns about subsistence resources to wildlife. The ethical concern therefore also lives with the ecological and livelihood risks associated with the zinc host system. A study published in 2017 found that metal concentrations in mosses near the haul road decreased between 2001 and 2006 but remained elevated relative to background conditions.[15] Because germanium production is associated with this process, ethical assessments must take these ecological effects into account.

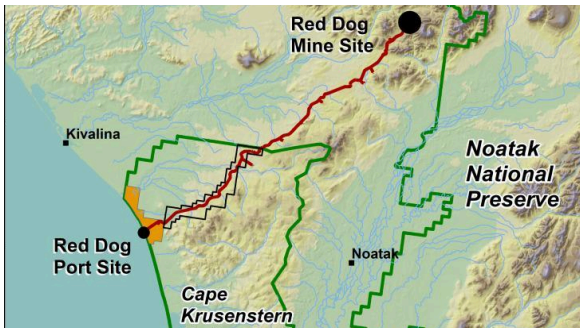


Figure 5. The Red Dog mine and DeLong Mountain Transportation System in relation to Cape Krusenstern National Monument and nearby communities. Neitlich et al., PLOS ONE 17(6): e0269801 (2022), Fig. 1 (CC0).

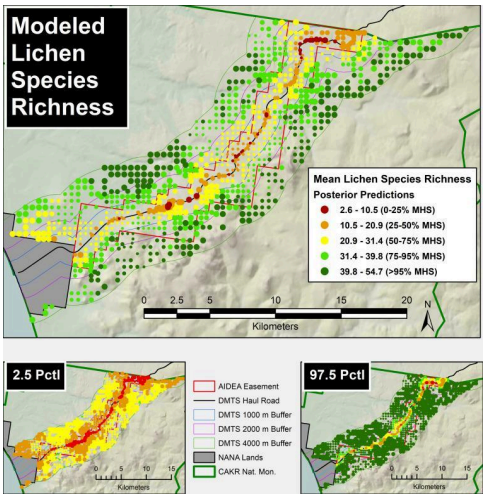


Figure 6. Mean lichen species richness in a prediction grid 0 to 4,000 m from the DeLong Mountain Transportation System (DMTS) haul road. Neitlich et al., PLOS ONE 17(6): e0269801 (2022), Fig. 7 (CC0).

4.3 Local Health Impacts and Remediation in Trail

The trail in British Columbia shows humanitarian impacts within the host-system footprint in the recovery of germanium. At Trail, germanium was produced[12] within the same complex responsible for more than a century of lead and zinc processing. Researchers investigated community exposure associated with contamination in surrounding soil and household dust.

A study published in 1991 reported average blood-lead concentrations of approximately 13.9 µg/dL in preschool children at the end of the 1980s.[16] The CDC observed that this was comparatively higher than the national average of 3.5 µg/dL.[17]

For many families, this contamination affected ordinary living spaces in their communities. In response to this legacy contamination, soil-management programs have tested residential yards and replaced contaminated soil where remediation was indicated, with special attention to households with young children.[18] This long-term response also includes voluntary annual childhood blood testing.[18] Among participating children aged 6 to 36 months in Trail County, the mean blood lead concentration fell from approximately 13.5 µg/dL in 1991 to 2.3 µg/dL in 2025.[18] Here, the ethical assessment is that host-system responsibility requires programs that prevent exposure and remediation programs to affected communities.

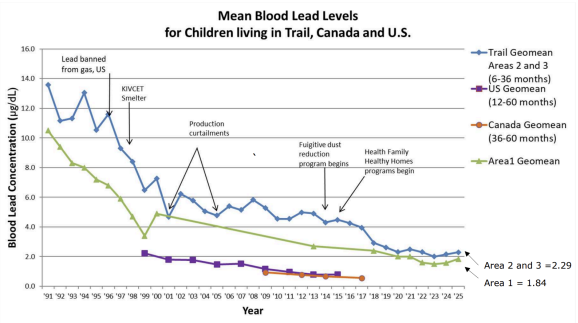


Figure 7. Mean blood-lead concentrations in Trail children, 1991-2025, against U.S. averages. Data: Trail Area Health & Environment Program, Fall 2025.[18,25]

4.4 Cross-Border Responsibility in the Upper Columbia River

Where the Trail case concerns exposure within the host community, the Upper Columbia River case looks at responsibility when host-system pollution crosses a national border. In December 2024, the EPA added the Upper Columbia River site in northeastern Washington to the Superfund National Priorities List.[19] The agency identified risks from lead and arsenic in residential soil, particularly for children and women, as well as risks to fish and wildlife from metals in river sediments.[19]

The EPA identifies the Trail smelter in Canada as the primary historical contamination source, but at the same time, does not exclude American sources of waste. For example, the EPA identified contributes from the Le Roi smelter in Washington.[19,20] EPA's June 2026 site update reported cleanup of contaminated soil at 87 residential properties.[20]

The ethical obligation to address these harms should follow effects across borders, rather than end at the jurisdiction in which the polluting facility operates.

As in Trail, the evidence is mostly about historical host-system contamination rather than harms demonstrated to originate specifically from germanium recovery. However, the Upper Columbia case shows how consequences and

responsibilities should extend across borders and generations. A proportionate response must include coordinated monitoring and remediation across regions, with meaningful participation by affected communities and governments.

5. Responsibility and Required Actions

5.1 Who is Responsible?

Assigning gallium or germanium an exact number of upstream environmental harm risks either understating or overstating their true impact. We emphasize transparent reporting of the direct recovery footprint and associated host-system impacts, without treating uncertainty over true responsibility as a reason to delay proactive action.

The case studies demonstrate that the environmental damage can become a humanitarian issue when it interferes with the resources people depend on. Across the cases, we observed consequences like deaths, injuries, and ecological damage. Studying numbers alone, like carbon emissions and waste quantities, cannot fully capture the dimensions of these harms.

Responsibility means determining which actors had the ability to cause, prevent, control, or respond to the harms associated with recovery operations and their host systems. The company operating a mine, for example, should bear responsibility since it controls the activity producing the environmental impact. Companies farther downstream may not have caused the original pollution, but they should investigate their sourcing relationships and use their purchasing leverage to encourage both proactive prevention and retrospective remediation.

We evaluate responsibility using three criteria: causation, control, and leverage. The Organisation for Economic Co-operation and Development distinguishes between companies that cause an adverse impact, contribute to it, or are directly linked to it through a business relationship.[21] We use these distinctions to prioritize obligations according to each actor's involvement in harm and capacity to prevent such cases from happening.

5.2 Mine and Refinery Operators

The strongest responsibility lies with the companies whose operations directly produce the documented environmental and humanitarian harms. Activities such as land clearing and residue storage can directly cause displacement or create pollution. Operators' agency over these activities gives them a direct capacity and corresponding responsibility to prevent and address harm. Where operators cause harm, they should prevent its continuation and cooperate in remedying for these harms.

These companies also often have greater access to information than surrounding communities. Operators should disclose relevant data on emissions and environmental risks to governments and affected communities. Access to this information places operators in a strong position to identify and address hazards arising from their processes. At Red Dog, Teck's operational role in mining and concentrate transport gave it direct influence and surveying power over the activities associated with dust dispersal.[12,14] Thus, Teck was able to adequately address contamination attributable to those activities. In other cases, operations that cause farmland loss should compensate and help to restore the land's continuing livelihood. Operators should fund environmental monitoring and community health organizations.[18]

In summary, direct causation and control place responsible operators first in the hierarchy of responsibility.

5.3 Governments and Regulators

Governments control the legal conditions under which mining and refining occur. Whether or not they hold ownership interests in an operation, governments have both

the authority and responsibility to establish ethical and environmental requirements. This can be done through incentives for improved practices, or stronger regulation policy for refineries.

This responsibility is important when communities have less bargaining power than mining companies. Inadequate enforcement from government can leave communities exposed to serious social and ecological harm. Governments have the duty to ensure that environmental standards protect the public welfare. Governments should require accessible environmental reporting and responsive when presented with unwarranted harms.

In its 2018 Guinea report, Human Rights Watch identified gaps in protections for customary land rights and in compensation for land losses.[10] The Guinean government also owns 49% of CBG, creating a potential conflict between its role as regulator and financial interest in mining.[22] Governments should separate regulatory decisions from their commercial interests and enforce protections against unfair mining costs to communities.

Finally, governments should require operators to secure funds for remediation and long-term waste management, even after closure. Ajka illustrates the reason for this decision; waste-storage safety should have the same regulatory attention as production, including maintenance and oversight after material recovery has ended.[8,9] Where contamination crosses borders, the relevant governments should coordinate a collective response rather than allow jurisdictions to leave communities affected.

5.4 Owners and Financiers

Owners and financiers might have operational control or influence through financing relationships, depending on their role. When available, they should use their shareholder voting and financing powers, to require credible prevention and remediation plans for affected communities. Their obligations should be proportionate to their involvement in the harm and their ability to influence the operator, instead of simply being dismissed solely because they do not manage day to day operations.

5.5 Downstream Buyers of Gallium and Germanium

Semiconductor and telecommunications companies may not directly cause upstream pollution, but just like financiers, their responsibilities depend on how their activities and purchasing relationships connect them to the harm. These companies should assess sourcing risks proactively and require suppliers to provide available traceability information and environmental practices.

Corporate reporting has historically provided examples of efforts to extend mineral-sourcing due diligence beyond traditionally monitored materials. Intel reported adding gallium and germanium to its mineral-sourcing due-diligence efforts in 2024.[23] In 2025, Intel sent the Responsible Minerals Initiative's Additional Minerals Reporting Template (AMRT) to suppliers contributing gallium, germanium, and several other materials to Intel-manufactured microprocessors. Approximately 68% of those suppliers responded to the survey, and Intel reported that it was continuing to identify refiners in these large supply chains.[23]

Downstream buyers should use these risks to require supplier improvements. Purchasing leverage can help to strengthen accountability when commercial expectations are tied to just demonstrable remediation. Buyers should publish accessible summaries of identified risks, and provide affected communities with opportunities to influence those responses.

6. Five-Point Plan for Sustainability

Goal: To establish an environmentally sustainable and socially responsible supply chain for gallium and germanium.

1. Implement transparent environmental reporting (Control). Require mining and refining companies to publish annual assessments of gallium and germanium recovery, including energy consumption and chemical waste generation reports.
2. Reduce pollution by improving waste management (Control). Require operators to monitor soil and groundwater surrounding the refineries, and governments to establish dedicated remediation funds to guarantee clean environments even after refineries close.
3. Protect local communities and livelihoods (Causation). Ensure that communities affected by mining receive access to independent environmental monitoring and health services, with residents actively participating in local policy.
4. Promote cleaner recovery technologies (Control). Invest in processes that reduce energy consumption and chemical waste during materials extraction such as flash Joule heating.
5. Establish government oversight (Leverage). Require semiconductor companies to trace their gallium and germanium suppliers and evaluate both environmental and humanitarian risks through governmental regulations.

7. Conclusion

Gallium and germanium are essential to modern technology, but their status as companion metals can hide the environmental and humanitarian impacts of the much larger industries on which they depend. The host-system impacts examined demonstrate long-term land loss, ecosystem damage, community exposure, as illustrated by the cases in rural areas such as Guinea and Alaska.

Responsibility should follow causation, control, and leverage. Mine and refinery operators should prevent continuing harm and remediate damage attributable to their operations. The clearest action is an enforceable requirement for operators to secure funds for remediation and long-term waste management that remain available after closure. Governments should enforce these obligations, while owners and financiers should use their influence to require.

Uncertainty about the share of host-system harm attributable to gallium or germanium should not excuse inaction by actors that cause, control, or have leverage over those harms.

8. References

[1]U.S. Geological Survey. Gallium: Mineral Commodity Summaries 2026. U.S. Geological Survey, 2026.

[2]U.S. Geological Survey. Germanium: Mineral Commodity Summaries 2026. U.S. Geological Survey, 2026.

[3]Svobodova-Sedlackova, Adela, et al. "Mapping the Research Landscape of Bauxite By-Products (Red Mud): An Evolutionary Perspective from 1995 to 2022." Heliyon, vol. 10, no. 3, e24943, 2024.

[4]U.S. Geological Survey. 2025 List of Critical Minerals. U.S. Geological Survey, 2025.

[5]Nassar, N. T., T. E. Graedel, and E. M. Harper. "By-Product Metals Are Technologically Essential but Have Problematic Supply." Science Advances, vol. 1, no. 3, e1400180, 2015.

[6]Luo, Hao, Tai-Yuan Huang, Xiaohan Wu, et al. "Life Cycle Assessment for Primary Gallium Production at Industrial-Scale." The International Journal of Life Cycle Assessment, vol. 30, pp. 1545–1559, 2025.

[7]Srivastava, Rajiv Ranjan, and Sadia Ilyas. "Harnessing Germanium from Industrial Residues and Electronic Waste for a Sustainable Energy Future." Green Chemistry, vol. 27, pp. 14744–14768, 2025.
[8]Government of Hungary, Ministry of Public Administration and Justice. "One Year On: Hungary Remembers the Red Sludge Disaster." Government of Hungary, 2011.
[9]Ruyters, Stefan, Jelle Mertens, Elvira Vassilieva, Boris Dehandschutter, André Poffijn, and Erik Smolders. "The Red Mud Accident in Ajka (Hungary): Plant Toxicity and Trace Metal Bioavailability in Red Mud Contaminated Soil." Environmental Science & Technology, vol. 45, no. 4, pp. 1616–1622, 2011.
[10]Human Rights Watch. "What Do We Get Out of It?": The Human Rights Impact of Bauxite Mining in Guinea. Human Rights Watch, 2018.
[11]Human Rights Watch. Aluminum: The Car Industry's Blind Spot—Why Car Companies Should Address the Human Rights Impact of Aluminum Production. Human Rights Watch, 2021.
[12]Teck Resources Limited. "Teck and DLT Partner to Pilot Traceability for Critical Minerals with Blockchain." Teck Resources Limited, 2022.
[13]Brumbaugh, William G., Suzette A. Morman, and Thomas W. May. "Concentrations and Bioaccessibility of Metals in Vegetation and Dust Near a Mining Haul Road, Cape Krusenstern National Monument, Alaska." Environmental Monitoring and Assessment, vol. 182, no. 1–4, pp. 325–340, 2011. DOI: 10.1007/s10661-011-1879-z.
[14]Neitlich, Peter N., Shanti Berryman, Linda H. Geiser, Anaka Mines, and Alyssa E. Shiel. "Impacts on Tundra Vegetation from Heavy Metal-Enriched Fugitive Dust on National Park Service Lands Along the Red Dog Mine Haul Road, Alaska." PLOS ONE, vol. 17, no. 6, e0269801, 2022.
[15]Neitlich, Peter N., et al. "Trends in Spatial Patterns of Heavy Metal Deposition on National Park Service Lands Along the Red Dog Mine Haul Road, Alaska, 2001–2006." PLOS ONE, vol. 12, no. 5, e0177936, 2017. DOI: 10.1371/journal.pone.0177936.
[16]Hertzman, C., H. Ward, N. Ames, S. Kelly, and C. Yates. "Childhood Lead Exposure in Trail Revisited." Canadian Journal of Public Health, vol. 82, no. 6, pp. 385–391, 1991.
[17]Centers for Disease Control and Prevention. "CDC Updates Blood Lead Reference Value." Centers for Disease Control and Prevention, 2024.
[18]Trail Area Health & Environment Committee. "2025 Children's Blood Lead Levels Remain Similar to Recent Years, with Participation Rates Rising." Trail Area Health & Environment Program, 2025.
[19]U.S. Environmental Protection Agency. "EPA Adds Upper Columbia River Site to the Superfund National Priorities List to Protect Public Health from Metals Contamination." U.S. Environmental Protection Agency, 2024.
[20]U.S. Environmental Protection Agency. Upper Columbia River Superfund Site. U.S. Environmental Protection Agency, updated June 5, 2026.
[21]Organisation for Economic Co-operation and Development. OECD Due Diligence Guidance for Responsible Business Conduct. OECD Publishing, 2018.
[22]Rio Tinto. CBG Sangaredi. Rio Tinto, accessed August 2026.
[23]Intel Corporation. Intel Responsible Minerals Sourcing Initiative—2025 Reporting Year. U.S. Securities and Exchange Commission, 2026.
[24]Microsoft Corporation. Conflict Minerals Report for the 2025 Reporting Year. U.S. Securities and Exchange Commission, 2026.
[25]Goodison, Karin, and Cecilee Pitman. Children's Blood Lead Levels in Trail: Fall 2025 Presentation. Interior Health, 2025.

Journal of Ethics in Scientific and Technological Innovations

Human-Centered Labor in the Age of AI: A Conceptual Analysis of Social and Emotional Work Structures



Isaac Sedighi[1]
[1]*Syosset High School, Syosset, NY 11791*

Abstract

Artificial intelligence is increasingly transforming the labor market by automating routine tasks, supporting decision-making, and changing service delivery. This conceptual analysis examines whether AI is likely to replace human workers in industries that depend heavily on social interaction, emotional intelligence, judgment, adaptability, and trust. It focuses on elementary education and the restaurant industry as two examples of human-centered labor. Research on elementary education suggests that AI is most effective when used as a digital scaffold that supports teachers rather than replacing them.[1,2] AI can help personalize instruction, improve assessment, and make learning more engaging, but it cannot authentically reproduce mentorship, empathy, creativity, or the relationships that shape a child's educational experience.[1,2] Similar limitations appear in one restaurant-service example: although robots may improve efficiency or attract customer interest, the Nairobi case reports continued preference for human service and current robotic systems as limited in function and expensive to implement.[3] This paper also considers the opposing argument that AI will create widespread unemployment due to its ability to reduce certain forms of human error and lower costs in various business endeavors. Recent layoffs demonstrate that AI has already disrupted some routine and technology-related jobs.[5] However, broader employment forecasts that combine technological, economic, demographic, and other macrotrends suggest that new roles may exceed those eliminated, resulting in a net increase in jobs by 2030.[6] The evidence therefore supports the interpretation that AI is transforming employment rather than ending it. Responsible implementation, human oversight, and possible labor protections may still be necessary to prevent excessive displacement. Overall, in the two cases examined, AI is more likely to serve as a supportive tool that enhances human performance while preserving the social and emotional foundations of human-centered work.

1. Introduction

Artificial intelligence is currently transforming the labor market and will continue to do so. It has the ability to automate routine tasks and reduce certain forms of human error in a multitude of different ways. As these capabilities develop, it is a major concern that the rise of AI will lead to a widespread increase in unemployment and the replacement of human workers.

However, this concern does not affect every part of the labor market equally. Interpersonal roles in industries such as education and hospitality remain essential because they rely on human judgment, emotional intelligence, adaptability, and trust. These characteristics are especially important in occupations in which social and emotional interaction is part of the service being provided. In this paper, AI refers primarily to

software-based systems, including generative AI, while robotics refers to physical systems that may incorporate AI. Human-centered labor refers to work in which interpersonal judgment, trust, and social or emotional interaction are integral to the service being provided.

This paper examines elementary education and the restaurant industry as examples of human interaction-centered labor. In both fields, AI may improve efficiency and support human workers but cannot fully reproduce the relationships, judgment, and authentic social interactions that these services require.[1,2,4] The evidence reviewed suggests that AI will transform the labor market by replacing some tasks and routine positions while preserving, and potentially increasing, the need for human-centered work. Elementary education and restaurant services are used here as illustrative cases in a conceptual discussion based on selected literature rather than as a comprehensive sample of the labor market.

2. AI as a Supportive Tool in Elementary Education

Fundamentally, education, specifically elementary education, is a human interaction-centered process. An effective elementary school teacher needs to develop a love of learning among the classroom's students, not just pass on information. According to one source, "the most effective education models use AI to support teachers, not replace them" because education relies heavily on "human judgment, emotional intelligence, creativity, and mentorship."[1]

A robot does not possess these abilities in the same way a human teacher does. A student may still perceive a robot or AI system as a role model, but that perception is distinct from whether the system forms a reciprocal human relationship. An AI system may learn patterns from interactions with students and generate responses that appear empathetic, but simulated empathetic behavior is not necessarily equivalent to human emotional understanding. This distinction is especially important in new and complex social or emotional situations involving young students.

Consequently, the role of AI in education will likely be supplemental rather than competitive. Building on this argument, a systematic review of AI in elementary education states that AI tools should be viewed as "digital scaffolds" that "supplement rather than supplant core learning experiences."[2] This concept places AI in a supporting role rather than positioning it as a replacement for the teacher or the educational experience itself.

Yim and Su's (2025) scoping review focuses primarily on AI learning tools and approaches for AI literacy in K–12 education, including pedagogical strategies, learning tools, assessment methods, and students' learning outcomes.[3] Its findings help describe how students learn about and with AI, but they do not by themselves establish that AI can personalize general instruction or assess student progress. Those broader applications require evidence from studies that examine them directly.

Education may need to adapt to the continued development of AI and other technologies. As students face increasing digital distractions and new opportunities to misuse generative AI for academic work, traditional, repetitive classroom models may become less effective at sustaining student attention and meaningful engagement.[7,8] As technology becomes more widespread, education must become more engaging and appealing to young students. By framing AI as a "scaffold," the research suggests that education will not disappear but will instead be strengthened.[2] The technology can help teachers give students the personalized learning they need and have often struggled to receive, while the actual "learning experience" remains rooted in human-led pedagogical principles.[2]

Overall, employment in education is not necessarily facing an inevitable downfall. Instead, education is undergoing a transformation in which it may survive and flourish by utilizing AI as a foundational support layer while maintaining essential human relationships at its original core.

3. AI and Human Service in the Restaurant Industry

A major part of the restaurant industry is the "human touch."[4] It is required in the dining room, kitchen, management office, and other parts of the dining environment. Again, while AI can improve efficiency in the industry, it cannot fully replicate the subtle but crucial social interactions associated with hospitality.

Edith Ojwang, a seasoned hospitality consultant, believes that it is possible for "robotic and human service to coexist" in the industry, but many clients continue to "prefer human service."[4] Her comments suggest that, in this Nairobi setting, some customers value human connection with waiters and other employees. Because this evidence comes from one reported example, it should not be generalized to customer preferences across the restaurant industry.[4]

The positive meaning associated with the term "homemade" provides a related but narrower example. Research on food authenticity found that traditional and homemade production practices can serve as cues consumers use when judging authenticity.[9] This evidence concerns food production and authenticity rather than staffing preferences, so it does not directly establish that customers prefer human restaurant staff.

According to a restaurant manager in Nairobi, robots are unable to "fully function in all the services" required in a dining environment and remain "too expensive" to be a viable cost-saving replacement for human labor.[4] In this specific setting, these observations illustrate two current limitations of AI and robotics in restaurant service: function and cost.

For AI to become the primary choice made by restaurant managers, it would need to function at or above the level of the contemporary restaurant worker while costing no more than the average human worker. At the moment, neither requirement is consistently met. However, partial automation can still reduce staffing needs even when a system cannot reproduce the full capabilities of a worker. A restaurant robot may perform specific tasks, attract customer attention, or serve as a novelty, but these limited functions do not equal the complete range of responsibilities handled by human restaurant employees.[4]

Therefore, full-service restaurants may continue to integrate AI and robotics as supplemental tools for efficiency or novelty while maintaining substantial human-led service. These conclusions apply most directly to settings where interpersonal service is central; faster-service or highly standardized restaurants may face different incentives for automation.

4. Labor-Market Displacement and Job Creation

The strongest opposing argument to the claim that AI will primarily support the human workforce rather than replace it is that AI is already eliminating jobs and may eventually cause widespread unemployment. Admittedly, employers cited generative AI in more than 10,000 job cuts during the first seven months of 2025.[5] Overall, more than 27,000 cuts in the technology sector have been attributed to AI since 2023.

These numbers demonstrate that the opposing argument is not misinformed. AI is already disrupting certain parts of the labor market, especially jobs involving routine, repetitive, or easily automated tasks. The immediate effects of this disruption can be

serious for workers whose current positions are reduced or eliminated.

However, this evidence does not prove that AI will cause a permanent and continuous loss of work across the economy as a whole. While layoffs are occurring, they demonstrate disruption in specific sectors without establishing the inevitable downfall of human employment.

The World Economic Forum projects that broad macrotrends—including technological change, the green transition, economic uncertainty, geoeconomic fragmentation, and demographic shifts—will create approximately 170 million jobs globally while displacing approximately 92 million, resulting in a net increase of 78 million jobs by 2030.[6] This projection means that the decline of some routine roles is not the full picture; those losses may be more than offset by the creation of new and specialized positions.

Some projected job growth is in technology, data, and AI roles, while other growing roles are expected in areas such as care, education, delivery, and agriculture.[6] Other roles will continue to require human interaction and judgment that cannot be easily automated.

Thus, the opposing argument correctly identifies a real short-term problem, but the available evidence does not establish that this disruption will become a permanent economy-wide decline in employment. A more cautious interpretation is that AI will transform the labor market by replacing certain tasks and routine positions while preserving the need for specialized and human-centered work.

5. Policy Responses and Responsible AI Integration

The innovators currently working to make AI as effective as possible should continue developing the technology. Based on the evidence reviewed, however, a machine is unlikely to fully replace the human role across the labor market. AI needs to be implemented responsibly and with proper oversight. With such oversight, it has greater potential to benefit society while limiting unnecessary harm to workers.

If extensive AI job replacement continues to disturb workers and the public, policy changes could be implemented. For example, governments could require certain businesses to ensure that a minimum percentage of total labor hours are performed by human employees rather than automated systems. A requirement could, for illustration, be discussed at a level such as 50%, although the specific percentage could be adjusted depending on the industry, national unemployment rate, and severity of worker displacement. This figure is illustrative rather than prescriptive; any real policy would need a clear method for measuring automated labor hours and hybrid human-AI tasks.

Another possible policy would be to establish tax penalties for excessive AI use or award tax benefits to businesses that maintain a high proportion of labor performed by human workers. These policies could encourage businesses to use AI to support their employees instead of immediately replacing them. Their purpose would be to preserve meaningful human roles and reduce abrupt displacement, although implementation would also need to account for compliance costs and the possibility of discouraging useful investment.

The magnitude of the situation should determine the type and strictness of the regulations created. Policies would need to be carefully managed to maintain economic stability without completely preventing technological development. Excessively strict regulations could limit useful innovation, while a lack of oversight could allow avoidable worker displacement.

Therefore, the objective of regulation should not be to stop AI development; it should

be to encourage responsible implementation, preserve necessary human employment, and ensure that businesses remain accountable for how automated systems affect their workers and customers.

6. Discussion

The evidence from education, hospitality, and broader labor-market forecasts supports a distinction between task replacement, staffing reduction, and complete job replacement. A system can reduce labor demand without reproducing every capability of a worker. Teachers and waiters, for instance, do much more than create slideshows and deliver food, respectively.

These findings suggest that the industries most resistant to full automation are those in which human interaction is not simply part of the production process but integral to the product or service itself. Students do not only receive information from teachers; they receive encouragement, mentorship, discipline, and emotional support. Restaurant customers do not only receive food; they receive attention, communication, and hospitality.

At the same time, current job losses demonstrate that the risks of AI should not be ignored.[5] Despite forecasts predicting a significant net increase in total employment across several macrotrends[6], the workers who lose existing jobs may not automatically qualify for the newly created positions. New specialized roles may require education, technical training, or experience that displaced workers do not currently possess. Therefore, a net increase in employment does not eliminate the need for retraining programs, worker protections, and responsible implementation.

The sources reviewed in this paper also have limitations. Employment projections are forecasts rather than guaranteed outcomes, and the long-term cost and capabilities of AI systems may change. Robots that are currently too expensive or limited[4] may become more affordable and functional. Nevertheless, the continuing importance of judgment, trust, empathy, and authentic social interaction suggests that human-centered labor will remain necessary even as the specific tasks performed by workers change.

7. Conclusion

AI is currently transforming the labor market and will continue to do so. It can automate routine tasks, reduce certain forms of human error, and improve efficiency in many industries. Current job cuts show that large-scale displacement is a legitimate concern in some sectors, especially for routine and easily automated positions.[5]

However, the available evidence does not establish that AI will cause the permanent downfall of human employment. One major forecast projects a significant net increase in total employment through 2030, but it attributes this change to several technological, economic, demographic, and environmental trends rather than to AI alone.[6] Human-centered occupations are also likely to remain very important because they depend on emotional intelligence, trust, adaptability, judgment, and genuine interpersonal connection.

As the technology matures, AI is most likely to serve in a supportive role that enhances human performance rather than uniformly competing with it.

Responsible implementation, human oversight, worker retraining, and carefully designed policies may still be required to manage the transition. When these protections are considered, AI has the potential to strengthen human work rather than eliminate it, ensuring that the social and emotional aspects of the labor market continue to play a central role.

8. References

[1]Mestri, Shadab. "Benefits of AI in Education: How It's Transforming Learning." The Education Magazine, https://www.theeducationmagazine.com/benefits-of-ai-in-education/.

[2]Boulhrir, Taoufik, and Mahmoud Hamash. "Unpacking Artificial Intelligence in Elementary Education: A Comprehensive Thematic Analysis Systematic Review." Computers and Education: Artificial Intelligence, vol. 9, 2025, article 100442. https://doi.org/10.1016/j.caeai.2025.100442.

[3]Yim, Iris Heung Yue, and Jiahong Su. "Artificial Intelligence (AI) Learning Tools in K–12 Education: A Scoping Review." Journal of Computers in Education, vol. 12, 2025, pp. 93–131. https://doi.org/10.1007/s40692-023-00304-9.

[4]Tiro, Desmond, and Evelyne Musambi. "Robot Waiters in Kenya Create a Buzz. But There Are Concerns about What It Means for Human Labor." AP News, 31 Aug. 2024, https://apnews.com/article/robot-cafe-waiters-kenya-technology-845cbb0bc610ef7f6fcbd86ef05803d2.

[5]Cerullo, Megan. "AI Is Leading to Thousands of Job Losses, Report Finds." CBS News, 5 Aug. 2025, https://www.cbsnews.com/news/ai-jobs-layoffs-us-2025/.

[6]World Economic Forum. "The Future of Jobs Report 2025." World Economic Forum, 7 Jan. 2025, https://www.weforum.org/publications/the-future-of-jobs-report-2025/.

[7]OECD. "Managing Screen Time: How to Protect and Equip Students Against Distraction." PISA in Focus, no. 124, OECD Publishing, 27 May 2024. https://doi.org/10.1787/7c225af4-en.

[8]Miao, Fengchun, and Wayne Holmes. Guidance for Generative AI in Education and Research. UNESCO, 2023. https://doi.org/10.54675/EWZM9535.

[9]Chousou, Charoula, and Konstadinos Mattas. "Assessing Consumer Attitudes and Perceptions towards Food Authenticity." British Food Journal, vol. 123, no. 5, 2021, pp. 1947–1961. https://doi.org/10.1108/BFJ-03-2019-0177.

Journal of Ethics in Scientific and Technological Innovations

The Limits of Algorithmic Authority: When Is Physician Override of Diagnostic AI Ethically Justified?

Steve Samuel[1]
[1]*Solon High School, Solon, OH 44139*

Abstract

As AI becomes increasingly prevalent, an ethical issue was created, specifically as to the role of physicians when AI systems demonstrate greater diagnostic accuracy than humans. This paper establishes the conditions a physician may ethically override a more accurate AI when diagnosing. Diagnostic accuracy is ethically important because errors can lead to inappropriate treatment, physical harm, financial burdens, and diminished trust between patients and healthcare providers. However, population-level diagnostic accuracy does not establish that an AI recommendation is correct for every individual patient. This paper compares the strengths and weaknesses of physicians and diagnostic AI to determine how their respective capabilities should influence clinical decision-making. Physicians contribute clinical reasoning, patient-specific information, and direct interaction with patients, but their judgment is also susceptible to diagnostic error, variability, fatigue, and overconfidence. AI can identify subtle patterns, provide consistent performance, and process information rapidly, but its performance may decline with atypical presentations, incomplete information, or unsupported outputs. These findings demonstrate that neither physicians nor AI should be assumed to be universally superior in every individual case. This paper proposes that initial deference should be given to AI when it has been proven to be more accurate in a diagnostic task, patient population, and clinical setting. However, the physician may override for a justifiable reason. This justification may come when patient-specific, evidence-based circumstances provide sufficient reason to believe that the AI recommendation is unreliable or inapplicable to the individual case. The strength of this justification should correspond to the AI's comparative reliability and the potential consequences of error. This framework preserves the ethical importance of diagnostic accuracy while maintaining a role for accountable physician judgment when individual circumstances warrant departure from AI recommendations.

1. Introduction

Accurate diagnosis is essential to patient care because diagnostic decisions determine the treatments and interventions that follow. When a diagnosis is incorrect, patients may experience inappropriate treatment, preventable harm, unnecessary financial costs, and diminished trust in healthcare providers. Diagnostic accuracy, therefore, has direct ethical significance because the consequences of diagnostic error can affect patient welfare.

Because AI has become more accurate, it also created a new problem for clinical decision-making. Diagnostic AI systems can now achieve diagnostic performance equal to, and sometimes greater, than that of physicians. For example, Hoppe et al. found that ChatGPT demonstrated significantly greater diagnostic performance than resident physicians across cardiovascular, endocrine, and gastrointestinal cases.[1] These developments create situations in which a physician may encounter an AI recommendation that conflicts with their own clinical judgment.

This creates an ethical problem that cannot be solved by determining whether physicians or AI are more accurate overall. This

is because a physician who disregards a more accurate AI system without proper justification may reject evidence that could improve the patient's diagnosis. However, treating AI accuracy as final in every case could prevent physicians from responding to patient-specific information that makes an AI recommendation less reliable. Thus, the central issue is when a physician has sufficient reason to stray from an AI recommendation that is generally more accurate.

Therefore the question is: under what conditions does physician override of a more accurate diagnostic AI remain ethically justified? This paper establishes that physician override is ethically justified when patient-specific, evidence-based considerations provide enough reason to believe that the AI's demonstrated reliability does not apply to the individual case. The strength of the justification should increase with the AI's comparative reliability and the clinical stakes of the decision.

To develop this argument, the paper first establishes the ethical importance of diagnostic accuracy and distinguishes population-level diagnostic performance from individual-patient reliability. Next, strengths and weaknesses of physicians and diagnostic AI before developing an ethical framework for justified physician override will be established. Finally, the practical implications, limitations, and future research associated with the framework will be considered.

2. Importance of Diagnostic Accuracy

This chapter aims to provide a workable definition of diagnostic accuracy that can be used in this paper. Once a definition is set, the significance of accuracy for patients must be established.

2.1 Definition of Diagnostic Accuracy

Šimundić describes diagnostic accuracy as "the ability of a test to discriminate between the target condition and health".[2] In medicine, diagnosis rarely depends on a single test. Instead, physicians usually use multiple types of evidence, which may include laboratory tests, imaging, patient history, physical examination, and other diagnostic assessments, which may depend on the condition being investigated. Therefore, diagnostic accuracy concerns the ability of a diagnostic method to correctly distinguish between patients with the condition of interest and those without it. The accuracy of a diagnostic test can be seen by comparing its results with an appropriate reference standard, to allow researchers to determine how frequently the test correctly identifies the presence or absence of a condition.

However, diagnostic accuracy can change due to different clinical circumstances. The variance may depend on "patient subgroups, with the spectrum of disease, with the clinical setting, with the test interpreters, and may depend on the results of prior testing".[3] Because the performance from a test depends on the circumstances where it was applied, evaluating the performance of a test requires consideration of the different circumstances in which the test was administered. This is especially important when considering the generalizability of a diagnostic method.

Therefore, for this paper, diagnostic accuracy is the ability of a diagnostic method to correctly distinguish the target medical condition from other conditions, while the level of accuracy is dependent on the population, clinical context, and information involved in the diagnostic process. This definition establishes that reported diagnostic accuracy must be interpreted with consideration to the conditions under which it was measured, something that will become important when comparing different diagnostic approaches.

2.2 Significance of Diagnostic Accuracy

Diagnostic error can cause problems to the patient, both emotionally and financially. An inaccurate diagnosis may lead patients to go through unnecessary treatment, putting them through costs without providing the intended medical benefit. The financial consequences of diagnostic error can also be substantial: a study analyzing U.S. malpractice claims from 1986 to 2010 found a mean inflation-adjusted payout of $386,849 per diagnostic-error claim.[4] This figure should not be interpreted as the average financial loss experienced by families, since malpractice payouts do not directly measure patients' total economic losses. Nevertheless, the magnitude of these payouts demonstrates that diagnostic errors can result in substantial compensable harm. Furthermore, unnecessary treatments expose patients to adverse effects, all while not treating the actual medical issue. A systematic review found that adverse drug reactions occurred in 8.32% of patients in primary care settings.[5] These physical and financial burdens can continue beyond the initial diagnostic error, particularly when the consequences of inappropriate treatment are not immediately recognized.

Diagnostic error can also affect how patients interact with the healthcare system. In a study examining patient experiences following ineffective treatment for depression, 23% of patients believed that their healthcare provider was not listening to their complaints, 28.9% reported sharing less information with their healthcare provider, and 14.5% reported missing their medications.[6] Although this study does not establish that diagnostic error caused these changes, ineffective treatment can result when a diagnosis is incorrect or insufficiently understood. The findings, nevertheless, demonstrate that ineffective treatment can alter how patients communicate with and respond to their healthcare providers. Because trust is a "fundamental component" of the physician-patient relationship,[7] changes in patient-provider communication following ineffective treatment could have important consequences for the patient's care in the future. Therefore, diagnostic accuracy is important not only because it provides the basis for selecting appropriate treatment, but also because accurate diagnosis may reduce the downstream harms associated with ineffective or inappropriate care.

Drawing on the previous two paragraphs, the inference can be made that the consequences of diagnostic error continue beyond the initial diagnosis of a patient's condition. Because a patient's diagnosis can affect the treatments that are given to them, there are downstream effects on the financial, physical, emotional, and relational aspects of their healthcare. Diagnostic errors have been associated with malpractice payouts averaging hundreds of thousands of dollars per claim, while inappropriate treatments may expose patients to adverse side effects, and ineffective treatment may affect the physician-patient relationship. Together, these consequences demonstrate that the importance of diagnostic accuracy extends beyond simply identifying a patient's condition correctly; it also affects the subsequent care that the patient receives. Therefore, diagnostic accuracy is an important component of patient welfare and quality clinical care.

2.3 Population Accuracy vs. Individual Accuracy

The importance of considering clinical context when evaluating diagnostic accuracy was established in Section 2.1. This dependence on context creates an important difference between the performance of a diagnostic method across a population and its reliability when applied to an individual patient. Population-level accuracy is how well a diagnostic method performs across the population of people in which it is evaluated, while individual-patient

reliability is how confidently the results of that method can be applied to a particular patient. Because diagnostic performance can change between populations, a measurement of accuracy that was taken from one population of patients cannot automatically be assumed to be useful for every individual patient. A diagnostic method may show strong performance across a study population while still giving an incorrect result for a particular patient. Therefore, the diagnosis of an individual requires the consideration of whether the evidence supporting that performance can be applied to the patient being evaluated.

Despite this distinction, population-level accuracy remains very relevant to clinical decision-making. Population-level accuracy allows clinicians to make inferences about how well a diagnostic method performs across a broader group of patients.[8] These measurements give an important foundation for determining whether a diagnostic method is generally reliable rather than relying solely on individual cases. However, because diagnostic performance may change when a method is applied to different populations, conclusions drawn from a single study may have limited generalizability in terms of application to different populations. For an individual patient, this can occur when characteristics such as age, comorbidities, disease severity, or clinical presentation may differ significantly compared to those represented in the study population. Differences in the clinical setting or in the type and quality of information available to the diagnostic method that is being used may also make the study population less representative of the patient being evaluated. This limitation can be addressed by assessing the diagnostic method across multiple populations and clinical conditions.[8] If similar levels of performance are seen across different studies and populations, clinicians can have greater confidence that the observed performance is not limited to the characteristics of a single study population. Nevertheless, even strong population-level evidence remains evidence about the performance of the method across groups of patients rather than a guarantee that its output will be correct for every individual.

The limitations of population-level accuracy become particularly visible when considering what common measures of diagnostic performance can establish about an individual patient. Akobeng explains that "Sensitivity and specificity are important measures of the diagnostic accuracy of a test but cannot be used to estimate the probability of disease in an individual patient".[9] Sensitivity and specificity give information about how a diagnostic test performs for patients who may or may not have the condition, but even these measures alone do not establish whether a specific patient has that condition. Consequently, even when sensitivity and specificity have been established through studies involving multiple populations, those measures cannot, by themselves, determine the probability that a particular patient has the disease. This distinction is important because a clinician does not ultimately need to know only how accurately a test performs across a population; the clinician must determine what the available evidence means for the patient being evaluated. Population-level measures can, therefore, inform an individual diagnostic decision without completely determining it.

The distinction between population-level accuracy and individual-patient reliability does not decrease the importance of diagnostic accuracy. Rather, it establishes the limits within which population-level accuracy should be interpreted. A diagnostic method that demonstrates high accuracy across multiple populations provides strong evidence that its results are generally reliable, and this evidence should influence confidence in its individual outputs. Individual diagnosis, therefore, requires

the population-level evidence to be considered with the particular circumstances of the patient to whom the diagnostic method is applied. This distinction becomes particularly important when two diagnostic approaches disagree. A difference in their overall accuracy provides evidence about their general reliability, but the circumstances of the individual patient may affect how applicable that evidence is to the particular diagnostic decision.

3. Accuracy of Physicians vs. AI

In order to better comprehend the reason for why AI is beginning to challenge physicians' authority in clinical settings, it becomes necessary to assess the level at which AI performs compared with physicians. Recently, there have been studies demonstrating that medical AI can have diagnostic performance similar to, or greater than, that of physicians. In a 2024 study, Hoppe et al. found that ChatGPT "demonstrated significant superiority" over resident physicians in cardiovascular, endocrine, and gastrointestinal diagnosis.[1] While the population used in the study had a median age of 72 years, these results clearly show that AI can achieve strong diagnostic performance across several clinically relevant areas. However, AI does not always exceed the performance of physicians across all measures of clinical performance. In a separate study, Robinson et al. asked non-medical volunteers to rate responses from chatbots and professional radiologists using a 5-point Likert scale, while subject matter experts independently rated the accuracy and completeness of the responses. The volunteers assessed the chatbot responses as being significantly more empathetic and expressed a greater preference for them, while experts found that chatbot responses were significantly more complete than those of the urologists. Nevertheless, there was no statistically significant difference in accuracy between the two groups, because both groups received average ratings corresponding to answers that were "nearly all correct".[10] Although this study does not demonstrate that chatbots are superior diagnosticians, it shows that diagnostic accuracy is not the only dimension in which AI and physicians may differ. To conclude, these studies demonstrate that neither physicians nor AI can be assumed to be superior diagnosticians. Rather, their respective strengths and weaknesses must be analyzed to determine the circumstances under which each may provide more reliable diagnostic information.

3.1 Physician Strengths

Physicians have several capabilities that can benefit their diagnostic accuracy. One important strength is their ability to integrate AI-generated information with their own clinical judgment. In a 2025 prospective, a multicenter study conducted across 67 medical organizations, researchers analyzed 3,409 brain CT studies for intracranial hemorrhage. Radiologists who used AI as an aid, outperformed the AI systems operating independently across all measured diagnostic metrics. Sensitivity increased from 95.91% for AI alone to 98.91% for AI-assisted radiologists, while specificity increased from 87.35% to 99.83%.[11] Because the AI systems were evaluated both independently and as tools for radiologists, these results provide evidence that physician involvement can add diagnostic value even when AI is already being used. This suggests that physicians can contribute judgment that improves upon the AI system's output on its own.

Another strength of physicians is their ability to use clinical reasoning to create and test diagnostic hypotheses. Hypothetico-deductive reasoning is a foundational component of clinical reasoning and is used by experienced clinicians when reasoning outside their typical intellectual domain.[12] This process involves generating possible diagnoses and testing them

against the patient's clinical information rather than only relying on immediate pattern recognition.[12] Physicians also develop "illness scripts," which organize knowledge about diseases according to their causes, pathological processes, symptoms, diagnostic findings, and expected progression.[13] These types of reasoning allow physicians to connect a patient's presentation to underlying disease mechanisms and think about other possible diagnoses when the available evidence does not fit an initial hypothesis. This provides a form of adaptive clinical reasoning that can build upon the pattern-recognition capabilities of diagnostic AI.

Physicians also obtain important information for diagnosing through direct interaction with patients. Clinical history is a major component of diagnosis, with some literature estimating that 80% of diagnoses can be made through history-taking alone in certain settings.[14] The physician-patient interview is not simply a fixed process of collecting predetermined information. Physicians can adjust their questions according to a patient's answers, check inconsistencies, and clarify vague symptoms.[14] Thus, the history may contain information that may not be available in an initial diagnostic image or dataset. Although AI systems are increasingly capable of conducting clinical conversation, physicians can utilize any information gathered during the patient encounter into their overall clinical assessment. This allows the physician to consider the patient's reported symptoms and circumstances alongside diagnostic tests and other clinical evidence.

Thus, the described strengths show that physicians can provide diagnostic expertise that extends beyond the analysis of one single dataset or image. Physicians can enhance the performance of an AI system, apply clinical reasoning to generate and test diagnostic hypotheses, and gather additional information through interaction with patients. These capabilities do not prove that physicians are inherently more accurate than AI. However, they demonstrate that diagnostic accuracy can depend on the integration of multiple forms of clinical information and reasoning. Physician judgment, therefore, remains a potentially valuable component of diagnosis, particularly when diagnostic information must be interpreted within the broader clinical context.

3.2 Physician Weaknesses

While physicians have their strengths, their diagnostic decision-making also has limitations that can affect patient care. First, diagnostic errors are relatively common. Researchers found that failures in diagnostic reasoning occurred in 66% of cases, diagnostic errors occurred in 13.8% of cases, and patients were harmed due to diagnostic errors in 11.3% of cases.[15] These findings demonstrate that errors can arise at multiple stages of the diagnostic process, from reasoning failures to errors that ultimately cause patient harm.

One source of this fallibility is the influence of human factors on diagnostic performance. Ivanovic et al. found that evening-time radiology diagnoses had a 33% higher error rate than morning diagnoses, while night-time diagnoses had a 53% higher error rate.[16] Although these findings do not directly establish fatigue as the cause of increased errors, they demonstrate that physician diagnostic performance can vary depending on the conditions under which a diagnosis is made. Physicians also face a trade-off between speed and accuracy. When physicians doubled their interpretation speed, their error rates increased by 270%.[16] This suggests that time pressure can negatively affect diagnostic accuracy, creating a limitation that is difficult to eliminate from human decision-making.

Physicians may also have difficulty recognizing their own diagnostic errors. Berner and Graber found that only 1% of physicians

reported having made a diagnostic error.[17] This is substantially lower than the prevalence of diagnostic errors reported in the literature, suggesting that physicians may be poorly calibrated regarding their own diagnostic performance. If physicians fail to recognize when their diagnoses are incorrect, they may be less likely to reconsider their reasoning or identify weaknesses in their decision-making process. This is particularly relevant when considering physician override of diagnostic AI, since confidence in a diagnosis does not necessarily indicate that the diagnosis is correct. Together, these limitations demonstrate that physician judgment is not only susceptible to error but can also vary according to circumstances and may not always accurately recognize its own mistakes.

Taken together, these limitations demonstrate that physician judgment is not consistently accurate or independent of human limitations. Physicians are susceptible to diagnostic errors; their performance can vary depending on factors such as time and workload, and they may not always recognize when their own diagnoses are incorrect. These weaknesses do not make physician judgment unreliable in every circumstance, nor do they eliminate the value of physician expertise. Rather, they demonstrate that physician judgment should not be assumed to be superior simply because it is human judgment. This distinction is particularly important when a physician considers overriding a diagnostic AI that has demonstrated greater accuracy, as the physician must have a justification beyond confidence in their own diagnosis.

3.3 AI Strengths

Now that the strengths and limitations of physicians have been considered, AI also has several strengths that may improve diagnostic decision-making. One of these strengths is its ability to identify patterns that humans may not necessarily recognize. McLeod et al. write that AI can identify "subtle features in diagnostic imaging scans that radiologists may miss, including higher-order features that lack obvious visual correlates".[18] This ability is demonstrated by research showing that AI models were able to determine a patient's sex from retinal scans with an accuracy of 87%.[18] The significance of this finding is that the AI was able to identify information from an image that was not previously thought to be obtainable from retinal scans. This demonstrates that AI can identify patterns within medical images that may not have an obvious visual correlation for human observers.

Furthermore, AI can demonstrate more consistent diagnostic performance than individual physicians. In a study by Hall, physicians demonstrated a sensitivity ranging from 54.2% to 80.7%, while the AI system demonstrated sensitivities of 87.4% and 77.7% under the reported testing conditions.[19] The variation among physicians demonstrates that individual physicians may differ substantially in their ability to correctly identify disease. In comparison, the AI's performance varied less across the reported conditions, suggesting that AI can provide a more consistent level of diagnostic performance. This consistency is particularly valuable in medical diagnosis because the accuracy of a diagnosis should not depend unnecessarily on which physician interprets the same information.

AI also has a substantial advantage in the speed and volume at which it can process medical information. When comparing AI and physicians in detecting intracranial large vessel occlusions (LVOs) on conventional CT angiography (CTA), AI reached a diagnostic decision in a mean time of 60 seconds, compared with an average of 395 seconds for physicians.[20] This demonstrates that AI can process certain diagnostic tasks substantially faster than physicians. AI can also process large

quantities of medical images without the same physical limitations that constrain human physicians. For example, one study reported that a radiologist in Korea analyzed approximately 936 images per month, which would amount to approximately 11,200 images per year.[21] The study described this as a substantial workload.[21] In contrast, another study found that AI could analyze a two-dimensional image in an average of 110 milliseconds.[22] At that rate, an AI system could theoretically process approximately 32,700 images in one hour. However, this figure should not be interpreted as the number of diagnoses an AI can make in an hour, since the 110-millisecond measurement refers to analysis of a two-dimensional image, whereas the 60-second measurement refers to analysis of a three-dimensional CTA and the production of a diagnostic decision. Nevertheless, these findings demonstrate the substantial difference between the processing capacity of AI and the practical limits of human physicians.

Taken together, these strengths demonstrate that AI can address several limitations inherent in human diagnostic decision-making. AI can identify subtle patterns that may be difficult for physicians to recognize, maintain more consistent diagnostic performance, and process large quantities of medical data at speeds that physicians cannot match. These capabilities can reduce some of the effects of human limitations such as variability, time pressure, and limited processing capacity. However, these advantages do not mean that AI should always replace physician judgment. Rather, they demonstrate that AI can provide diagnostic information that may be more accurate, consistent, or efficient than human judgment in certain circumstances. This gives AI an important role in diagnostic decision-making while leaving open the question of when physician judgment can appropriately take precedence.

3.4 AI Weaknesses

While AI has demonstrated strong diagnostic performance in many clinical cases, its performance may decline when dealing with atypical or nonspecific presentations of medical issues. Shikino et al. found that "ChatGPT's diagnostic accuracy decreased with an increase in atypical presentation".[14] This issue can also be seen in a study comparing AI and physicians across three levels of disease presentation specificity.[23] When the AI was presented with highly specific disease presentations, it achieved 100% accuracy. Its accuracy decreased to 83% for moderately specific presentations and 72% for nonspecific presentations.[23] Although the AI outperformed the physician in each of the three rounds, its decreasing accuracy as the presentations became less specific demonstrates that strong diagnostic performance on recognizable cases does not necessarily mean that the AI will have strong performance on less characteristic cases. This is particularly important to individual-patient diagnosis because a patient may not present symptoms in the manner most strongly associated with their underlying condition. Therefore, an AI system's overall diagnostic accuracy may provide strong evidence of its general reliability while still being less predictive of its performance in an atypical individual case.

A second limitation is about AI's capability to determine what information is necessary to reach a correct diagnosis. Hager et al. tested several AI models, including Llama 2 Chat, OASST, and WizardLM, and found that the models under-requested the necessary testing required to reach the correct diagnosis in the cases evaluated.[24] The models also underprescribed treatments, especially for patients suffering from more severe pathologies.[24] Therefore, the efficiency of AI models' diagnoses depends not only on its ability to interpret information that it receives, but also on its ability to recognize what

additional information is necessary to make an appropriate decision. It is a key feature because a diagnostic system may provide a good result when working with sufficient information while still performing poorly at identifying needed information independently. For an individual patient, failing to request a relevant test or piece of information may prevent the system from recognizing a condition that does not fit the information it has already received. Thus, the limitation is not simply that AI can make an incorrect diagnosis, but that it may not identify the additional information necessary to avoid that incorrect diagnosis.

Finally, AI models can produce information that is not grounded in the available evidence. Hager et al. found that the tested AI models were "hallucinating nonexistent tools every two to five patients".[24] This presents a reliability problem because that means that AI's diagnosis may be influenced by information/resources that may not really exist. Unlike physicians, who may make flawed interpretations of real clinical evidence, an AI system that hallucinates nonexistent information can undermine the basis of its recommendation. If that hallucinated information contributes to an incorrect diagnosis or treatment recommendation, the resulting error could produce the physical, financial, and relational harms discussed in Section 2.2. Therefore, hallucination represents a limitation not only because it can produce incorrect information, but also because it can make an AI-generated recommendation appear to have evidentiary support when that support is nonexistent.

Collectively, all these shortcomings imply that the level of AI's diagnostic accuracy cannot be seen as an indicator of reliability of individual cases. AI can perform very well when it is faced with obvious cases, however it may fail when there is additional information (and not requested by AI) or when AI's logic is based on hallucinated information. All of these factors combined prove the point that the level of AI's diagnostic accuracy is not enough in order to estimate the reliability of each AI suggestion.

All things considered, the comparison between physicians and AI shows that while AI may outperform physicians in certain areas, it can still be wrong in specific cases. The physician is able to contribute to the process of clinical reasoning, information about the patient, and professional judgement, while AI has strengths when it comes to pattern recognition, consistency, processing speed, and large-scale data analysis. However, both entities are susceptible to making mistakes due to different reasons: physicians may do well or may not do well due to human limitations and circumstances, while AI may perform worse in atypical cases, depending on the information, or producing unsupported information. Therefore, the relevant question from here is not whether physicians or AI as a whole are more accurate, but whether the circumstances of a particular case provide sufficient reason for one to take precedence over the other.

4. Ethical Justification of Physician Override

Physician override of a superior diagnostic AI is ethically justified not merely when a physician disagrees, but when patient-specific evidence provides sufficient reason to believe that the AI's recommendation is unreliable in the individual case. The strength of this justification should depend on the AI's comparative reliability and the clinical stakes of the decision.

4.1 The Presumption of AI Deference

The importance of diagnostic accuracy established in Chapter 2 provides an ethical reason to give serious weight to an AI system that has demonstrated superior performance for a relevant diagnostic task. The consequences of diagnostic errors discussed in Chapter 2.2 demonstrate the importance of making accurate

diagnostic decisions. From the principle of beneficence, since physicians have a duty to promote patient welfare and avoid preventable harm,[25] they should not disregard a diagnostic method that is more likely to identify the patient's condition correctly without a sufficient reason to do so. Thus, when an AI system has been reliably shown to outperform physicians on the relevant diagnostic task and patient population, its recommendation creates an initial presumption in favor of deference.

This presumption does not mean that AI should replace physician judgment or that physicians have no role in evaluating an AI recommendation. Rather, it means that a physician's professional status, personal confidence, or mere disagreement with the system is not by itself enough to justify rejecting a more accurate diagnostic method. Chapter 3.2 established that physician judgment is susceptible to diagnostic error, time pressure, variation in performance, and poor recognition of its own mistakes. If a physician could override an AI system solely because they preferred their own diagnosis, the potential accuracy advantages of AI could be ignored for reasons unrelated to the patient's best interest. Such a practice would risk treating human authority as more important than the reliability of the available evidence.

The presumption of deference is especially important because AI can identify patterns that physicians may miss, produce more consistent results across some diagnostic tasks, and process medical information at speeds beyond normal human capacity. These advantages do not establish that AI is always correct, but they do establish that its recommendation may have significant evidential value. When a system has demonstrated superior accuracy for the relevant task and population, following its recommendation is ordinarily more likely to support an accurate diagnosis than disregarding it without further inquiry. A physician, therefore, should take the recommendation seriously and investigate a meaningful disagreement rather than simply dismissing it.

However, the presumption in favor of AI deference must remain rebuttable. As Chapter 2.3 established, population-level accuracy is not identical to reliability in an individual patient's case. A diagnostic system may perform strongly across a population while still being less reliable when applied to a patient whose presentation, history, or circumstances differ from those represented in the evidence supporting the system's reported accuracy. As discussed in Chapter 3.4, AI performance may decline in atypical cases, may depend on whether it receives adequate information, and may be undermined by hallucinated or unsupported content. Therefore, superior aggregate performance gives an AI an initial claim to deference, but it does not establish that its output must always control the decision in a particular case.

The appropriate ethical position is consequently neither blind AI deference nor unrestricted physician authority. When AI is demonstrably more reliable for the relevant diagnostic task, physicians should begin with a rebuttable presumption that its recommendation deserves to be followed. To override that recommendation ethically, however, the physician must be able to identify a patient-specific, evidence-based reason to believe that the AI's general reliability does not apply to the individual case.

4.2 Patient-Specific Context and Clinical Judgment

While a more accurate AI system initially prompts a stance of deference, the presumption depends on whether the evidence that supports the system's reported accuracy actually applies to the patient being evaluated. As established in Chapter 2.3, while population

accuracy describes how a diagnostic method performs across a group of patients, it does not establish that the method's output is correct for every individual. The ethical importance of this difference is that physicians should not defer to AI merely because the system works well on average. Rather, the physician should take into account whether the individual patient's circumstances gives reason to believe that the AI's general performance is not reliable in the situation. However, patient-specific context is relevant in many ways. First, a physician may have information that was not available to the AI system when it gave its recommendation. Diagnostic AI often operates on a defined set of inputs, such as an image, laboratory value, symptom list, or electronic health record. A physician, by contrast, may obtain additional information through direct examination, conversation with the patient, review of the patient's medical history, or observation of changes in the patient's condition. As discussed in Chapter 3.1, clinical history is a very important component of diagnosis, and physicians can adapt their questions when a patient's responses are ambiguous or inconsistent. This additional information can give additional evidence that was not available to the AI when it generates its recommendation.

Second, the patient may have an unusual presentation of a medical condition that makes the AI's general accuracy less applicable to that situation. Chapter 3.4 established that AI diagnostic performance can worsen as disease presentations become less specific or less characteristic of a disease. This is ethically important because even if an AI demonstrates strong performance in typical cases, it does not justify the same level of confidence when the patient's situation differs substantially from the cases on which the system was taught. For example, a patient may have an unusual combination of symptoms, an uncommon comorbidity, or a presentation altered by age, pregnancy, medication, or an underlying condition. In such situations, the physician's observation that the case is atypical is not just an expression of disagreement, it may be evidence that the conditions supporting the AI's reported accuracy are absent.

Clinical judgment can also provide a reason to question an AI diagnosis recommendation when the recommendation does not fit the patient's broader clinical picture. As described in Chapter 3.1, physicians use hypothetico-deductive reasoning to create possible diagnoses and test them against the evidence that is available. They also draw on illness scripts that connect symptoms, causes, pathological processes, diagnostic findings, and expected disease progression.[13] However, as discussed earlier, these methods do not make physician judgment automatically superior to AI. Indeed, Chapter 3 demonstrated that physicians are prone to error and may be overconfident in their own diagnoses. Nevertheless, clinical reasoning can identify a meaningful inconsistency between an AI output and the patient's history, examination, test results, or observed progression. When the physician can explain that inconsistency through relevant evidence, the disagreement may weaken the case for deference to AI.

The difference between justified judgment and unjustified intuition needs to be understood. A physician does not have ethical grounds to override a more accurate AI merely by asserting that the recommendation "does not feel right." Because physician reasoning can be influenced by bias, fatigue, and diagnostic error, an override should be grounded in identifiable features of the case. Any physician that seeks to overturn an AI's decision needs to point out what information was lacking in the case, what was unusual about the presentation of the issue, or why the recommendation is contradictory to other evidence. This requirement protects

against the risk that professional confidence will be mistaken for patient-specific justification.

The patient-specific context may serve as a rebuttal to the presumption of AI deference if it provides convincing proof that the performance of AI in general populations is not reliable enough to be transposed onto the specific patient case. The role of the physician from the ethical standpoint is not simply to substitute personal judgment for an AI recommendation. Rather, it is to determine whether the patient's particular circumstances give reason to revise the confidence that should be placed in that recommendation.

4.3 Uncertainty and Limits of AI Accuracy

The presumption that the physician should defer to AI should not be understood as a rule that the AI is always correct. Even when the AI has shown higher diagnostic accuracy than physicians, the conclusion concerns how often the system produces correct results across the relevant population, not other populations. Furthermore, it does not establish that the system's recommendation is always correct in every individual case. As Chapter 2.3 explained, diagnostic performance measures like sensitivity and specificity give evidence about reliability of a method across groups, but they cannot by themselves give enough evidence for a particular patient's diagnosis. Thus, the ethical weight given to an AI diagnosis should reflect its reliability that it demonstrated without treating it as a certainty. As a hypothetical scenario, if an AI system is accurate 95% of the time under comparable conditions, it may still produce an incorrect result in about 5% of cases. While this is a strong diagnostic performance (especially if it exceeds physician performance), it does not make the system infallible. A physician who gives evidence that suggests that a particular patient may fall outside the system's accuracy ability is not rejecting the importance of accuracy. Rather, the physician is evaluating whether the system's general reliability provides sufficient reason to control the decision in this individual case.

The reliability of an AI recommendation may also depend on what information is given to it, and the quality and completeness of the information. As discussed in Chapter 3.4 AI may struggle when it needs to identify what additional testing or information is needed to reach an accurate diagnosis. Thus, a diagnosis from AI may be less reliable when the system did not receive an important test result, an important detail from the patient's history, or necessary evidence of a change in the patient's condition. Similarly, some AI systems may hallucinate information, meaning an explanation or piece of evidence given by the system may not give genuine support for its recommendation. In order to determine if there is sufficient information present for the AI and if the reasoning or evidence is consistent with the current clinical data, the amount of information and the quality of the information given to the AI should be evaluated by the physician. A detailed or confident explanation should not automatically be treated as evidence that the underlying recommendation is reliable.

The appropriate response to uncertainty should also depend on the potential consequences of the decision. A small possibility that an AI recommendation is incorrect may bring the need for additional investigation when following the diagnosis could lead to severe or irreversible harm. However, the need for additional investigation does not by itself establish that the physician should override the AI. Where the consequences of an incorrect decision are minor or readily reversible, the same degree of uncertainty may not justify departing from a more accurate diagnostic system. Physicians should therefore consider both the reliability of the AI recommendation and the potential consequences of following or overriding it. The greater the

AI's demonstrated reliability in that field and the stronger the evidence supporting its recommendation for its diagnosis, the stronger the justification the physician must provide for override. Conversely, when meaningful uncertainty combines with substantial potential harm, additional investigation or physician intervention may be ethically warranted to provide a higher probability of an accurate diagnosis. This relationship between evidential strength, uncertainty, and clinical consequences provides a basis for determining when physician override is ethically justified.

4.4 Physician Responsibility and Accountability

Physician override of a more accurate diagnostic AI carries a corresponding responsibility for the decision to depart from the system's recommendation. Because the AI has an established presumption of deference, a physician who chooses to override it is exercising discretionary authority against a method that has demonstrated superior performance for the relevant task. This does not mean that the physician must prove that the AI is wrong or that an incorrect override is automatically unethical. Rather, the physician should be able to explain why the decision to depart from the AI was reasonable given the evidence and circumstances available at the time.

This requirement distinguishes justified clinical judgment from an unreviewable claim of professional authority. A physician should not be able to justify an override solely by stating that the decision was based on "clinical judgment" or that the AI recommendation did not seem correct. Instead, the physician should be able to identify the basis for the decision, such as patient-specific evidence that conflicted with the AI recommendation, information unavailable to the system, an atypical presentation, or a limitation in the AI's available inputs. The purpose of this requirement is not to require certainty, but to ensure that physician authority is exercised on the basis of reasons that are relevant to the individual patient.

Accountability also requires that the reasoning behind an override be capable of review. When circumstances permit, documenting the basis for an override allows other clinicians to evaluate whether the decision was supported by the available evidence and provides an opportunity for consultation or reconsideration. This is particularly important when the physician's decision could expose the patient to significant harm. Documentation does not guarantee that the decision will be correct, but it makes the reasoning underlying the decision transparent rather than allowing professional status alone to determine whether an override was justified.

Importantly, this standard should be evaluated using the information available when the decision was made rather than the outcome that followed. A physician may make a well-supported decision to override an AI and still reach the wrong diagnosis. Conversely, an override should not become ethically justified merely because the physician happened to be correct. The relevant question is whether the physician had sufficient patient-specific reasons to depart from the AI given the evidence, uncertainty, and clinical circumstances at the time.

Therefore, physician responsibility does not require either unquestioning reliance on AI or proof that the AI is incorrect. It requires that physicians recognize the authority of a more accurate diagnostic system, exercise override authority only for reasons relevant to the individual case, and remain able to explain and defend that decision. This accountability requirement provides an additional safeguard against replacing evidence-based decision-making with either blind AI reliance or unsupported physician confidence. The final section can therefore combine the requirements

established throughout this chapter into specific conditions under which physician override is ethically justified.

4.5 Conditions for Justified Override

The preceding sections establish that neither AI's superior aggregate accuracy nor a physician's professional judgment should automatically control an individual diagnostic decision. When an AI system has demonstrated greater reliability for the relevant task, setting, and patient population, its recommendation should receive presumptive deference. However, this presumption can be rebutted when case-specific evidence provides sufficient reason to believe that the AI recommendation is unreliable or inapplicable to the individual patient. Physician override is therefore ethically justified only when the physician's reasons are evidence-based, responsive to the individual case, proportionate to the reliability of the AI and the consequences of the decision, and subject to appropriate accountability.

First, the physician must identify a patient-specific reason to question the applicability or reliability of the AI recommendation. This reason may arise because the AI lacked relevant patient information, because the patient's presentation is atypical, because the output conflicts with credible clinical findings, or because the system's recommendation appears to rest on incomplete or unsupported information. The relevant concern is not simply that the physician has reached a different conclusion. It is that there is an identifiable feature of the case that weakens the inference from the AI's population-level performance to its reliability for this patient.

Second, the physician's alternative judgment must be evidence-based. A physician should be able to explain why the available patient information supports an alternative diagnosis or course of action. This does not require proof that the AI is certainly incorrect; medicine often requires decisions under uncertainty. It does, however, require more than intuition, confidence, or a general belief that human clinical judgment should take priority. If the physician cannot identify credible evidence supporting the alternative judgment, disagreement alone does not justify an override.

Third, the strength of the justification required should correspond to the strength of the evidence supporting the AI's reliability in the relevant circumstances. An AI system with only a small or uncertain performance advantage creates a weaker presumption of deference than a system repeatedly shown to be substantially more accurate for the same task, setting, and patient population. The presumption should also be weaker when there is reason to believe that the evidence supporting the AI's reported performance does not generalize to the individual patient. Conversely, when the AI has demonstrated a substantial and well-supported advantage under circumstances closely matching the patient's case, stronger patient-specific evidence should be required before the physician rejects its recommendation.

Fourth, the decision to override should be proportionate to the clinical stakes. When following either the AI's or the physician's alternative judgment could result in severe or irreversible harm, uncertainty may justify additional investigation rather than an immediate choice between the two recommendations. Further testing, consultation, monitoring, or referral may be ethically preferable when these measures can clarify the disagreement without exposing the patient to unnecessary risk. Conversely, if delay itself could cause serious harm, the physician may need to act on the best available evidence without complete certainty. The appropriate response to disagreement therefore depends not only on the relative reliability of the AI and physician, but also on the potential consequences of error, delay, and intervention.

Finally, an override should be accountable. When circumstances permit, the physician should document the patient-specific basis for departing from the AI recommendation and explain how the available evidence supports the alternative decision. This requirement does not treat documentation as a substitute for sound judgment. Rather, it ensures that the override can be evaluated as a reasoned clinical decision rather than an unexplained rejection of a more accurate diagnostic tool. In urgent circumstances, the explanation may need to follow the immediate decision, but the physician should still be able to identify the reasons that justified acting.

Physician override of a superior diagnostic AI is therefore ethically justified when five conditions are met: the physician identifies a patient-specific reason to question the AI's applicability or reliability; the alternative judgment is supported by credible evidence; the strength of the justification is proportionate to the AI's demonstrated reliability in the relevant circumstances; the decision is proportionate to the potential consequences of error, delay, and intervention; and the physician remains accountable for the decision. This standard preserves the ethical importance of diagnostic accuracy without treating population-level superiority as a substitute for judgment about the individual patient.

5. Discussion

5.1 Implications for Clinical Practice

The framework developed in this paper suggests that clinical disagreement between physicians and diagnostic AI should be treated as a decision point rather than as an automatic reason for either deference or override. When a more accurate AI system produces a recommendation that conflicts with a physician's judgment, the physician should evaluate whether the disagreement reflects a genuine limitation in the AI's applicability to the individual patient. This requires considering the information available to the system, the characteristics of the patient's presentation, and the other clinical evidence relevant to the diagnosis.

In practice, this framework would shift physician judgment away from functioning as an independent source of diagnostic authority and toward evaluating when AI's general reliability applies to the individual patient. A physician should therefore be able to identify the specific feature of a case that justifies departing from the AI recommendation. Relevant considerations may include an atypical presentation, information unavailable to the system, conflicting examination or test findings, or evidence that the AI's input was incomplete. This does not require physicians to prove that the AI is wrong; rather, it requires them to determine whether sufficient evidence exists to weaken the presumption that the AI's recommendation should be followed.

The framework also suggests that disagreement should not necessarily result in an immediate choice between the physician's diagnosis and the AI's diagnosis. When circumstances permit, disagreement can instead trigger additional diagnostic investigation. A physician might review the information provided to the AI, obtain additional testing, consult another clinician, or monitor the patient's condition before making a final decision. Such approaches are particularly relevant when the consequences of either diagnostic error are severe or irreversible. In these situations, the ethical objective is not simply to determine which source is more trustworthy, but to reduce uncertainty when doing so is reasonably possible.

The framework also has implications for accountability within clinical practice. When a physician overrides a more accurate AI system, the basis for that decision should be identifiable and, when feasible, documented. Documentation can distinguish an evidence-based override from

one based primarily on confidence or intuition and can allow other clinicians to evaluate the reasoning behind the decision. This is particularly important in complex cases where multiple clinicians may participate in patient care. In urgent situations, immediate action may prevent detailed evaluation or documentation beforehand, but the physician should still be able to identify the patient-specific circumstances that justified departing from the AI recommendation.

Overall, the framework supports a model of clinical practice in which diagnostic AI functions as a highly reliable source of evidence rather than either an unquestionable authority or a passive tool. Physicians remain responsible for determining whether that evidence applies to the individual patient. The practical implication is therefore not simply that physicians should "work with AI," but that disagreement between physician and AI should become a structured opportunity to evaluate applicability, investigate uncertainty, and justify departure from the more accurate system when the evidence warrants it.

5.2 Implications for Diagnostic AI Design and Deployment

The framework developed in this paper implies that diagnostic AI systems should provide clinicians with sufficient information to assess whether an output is reliable for the individual patient. A recommendation alone may be insufficient for this purpose. Where possible, clinicians should be able to determine the clinical task, patient population, and setting in which the system was evaluated, as well as the relevant information provided to the system when generating its recommendation. This information would allow physicians to better assess whether the evidence supporting the AI's reported accuracy applies to the patient before them.

AI systems should also communicate relevant uncertainty and limitations. Where technically feasible, systems should identify missing or insufficient inputs, indicate when a case falls outside validated conditions, and distinguish between recommendations supported by sufficient information and those generated under substantial uncertainty. Such information would help physicians determine whether the presumption of AI deference established by this paper should remain strong or whether patient-specific circumstances warrant further investigation.

System design should also support appropriate evaluation of the basis for an AI recommendation. Because AI systems can generate unsupported or erroneous information,[24] a detailed explanation should not automatically be treated as evidence that the recommendation is correct. Where reliable methods exist, AI systems should identify the clinical information, test results, or other relevant features that contributed to their output and communicate important limitations in that reasoning. The objective should not be to create the appearance of explainability, but to provide clinicians with information that can genuinely assist them in determining whether the recommendation is supported by the available evidence.

The framework also has implications for healthcare organizations deploying diagnostic AI. Policies should require meaningful consideration of AI recommendations without preventing physicians from overriding them when patient-specific evidence provides sufficient justification. When circumstances permit, documenting the basis for an override could support accountability and allow the decision to be reviewed by other clinicians. Such procedures should remain practical in time-sensitive settings and should not create delays that themselves endanger patients.

Ultimately, diagnostic AI should be designed and deployed to support accountable clinical decision-making rather than to establish

an unquestionable source of authority. The more accurately a system performs a relevant diagnostic task, the more important it becomes that clinicians can determine when that demonstrated accuracy applies and when it may not. AI design should therefore provide the information necessary for physicians to evaluate applicability, uncertainty, and limitations, allowing the system's diagnostic strengths to be used without treating its outputs as infallible.

5.3 Limitations of the Framework

This framework does not provide a precise numerical rule for determining when physician override is justified. It identifies the relevant considerations—patient-specific evidence, the AI's comparative reliability, and the clinical stakes—but determining whether those considerations are sufficient will still require clinical judgment. This reflects the fact that diagnostic decisions often must be made under uncertainty.

The framework also depends on establishing that an AI system is more accurate for the relevant diagnostic task, patient population, and clinical setting. In practice, that determination may be difficult to make. As established in Section 2.3, diagnostic performance can vary between populations and clinical circumstances, meaning that strong performance in one setting does not necessarily establish equivalent performance in another. A claim that AI is more accurate should therefore be evaluated carefully before it creates a strong presumption of deference.

Finally, this paper focuses on diagnosis rather than treatment. Treatment decisions involve patient values, preferences, quality of life, and acceptable levels of risk in ways that diagnostic accuracy alone cannot determine. The framework may inform those decisions, but further ethical analysis would be needed to apply it directly beyond diagnostic AI.

5.4 Future Research

Future research should examine how physicians can determine whether an individual patient falls outside the circumstances in which a diagnostic AI system has demonstrated reliable performance. Studies could investigate how atypical presentations, incomplete inputs, and variation across patient populations affect AI performance and whether clinicians can reliably recognize when these factors reduce the applicability of an AI recommendation. Such research could help make the framework's patient-specific standard more practically usable.

Research should also examine how AI systems communicate uncertainty and limitations to clinicians. Different methods of presenting confidence, missing information, and the basis for a recommendation could be compared to determine which approaches help physicians recognize meaningful limitations without encouraging either excessive trust or excessive skepticism. This research could inform the design of AI systems that better support the type of evaluation required by the framework.

Finally, empirical research should examine how physicians respond to disagreement with AI in actual clinical practice. Studies could evaluate the reasons physicians give for overriding AI recommendations, whether requiring documentation affects the quality of those decisions, and whether additional testing or consultation improves outcomes when physician and AI recommendations conflict. Research in these areas could determine whether the framework's proposed conditions for override are practical and whether they lead to better diagnostic decision-making in real clinical settings.

6. Conclusion

This paper examined whether physicians may ethically override a diagnostic AI system that has demonstrated greater accuracy than

physician judgment. Diagnostic accuracy is ethically significant because inaccurate diagnoses can lead to inappropriate treatment, physical harm, financial burdens, and diminished trust in healthcare providers. Therefore, when an AI system has demonstrated superior diagnostic performance, its recommendation cannot be dismissed merely because it conflicts with a physician's initial judgment.

However, superior population-level accuracy does not guarantee correctness in every individual case. Diagnostic AI can identify subtle patterns, provide consistent analysis, and process clinical information rapidly, but its performance may decline when patients present atypically, when relevant information is unavailable to the system, or when its recommendations rely on unsupported information. Physicians likewise possess important capabilities, including clinical reasoning and direct access to patient history and context, but their judgment is subject to diagnostic error, fatigue, bias, and overconfidence. Neither source of diagnostic information should therefore be treated as universally authoritative.

The appropriate ethical relationship between physicians and diagnostic AI is consequently neither unrestricted physician authority nor blind deference to AI. When an AI system has been reliably demonstrated to outperform physicians for the relevant diagnostic task, setting, and patient population, its recommendation should receive a rebuttable presumption of deference. Physicians should take such recommendations seriously and should not override them solely because of professional status, confidence, or unsupported disagreement. At the same time, population-level superiority does not eliminate the possibility that the AI is unreliable or inapplicable in an individual case.

Physician override is ethically justified when the physician identifies patient-specific, evidence-based reasons to question the AI recommendation; when the alternative judgment is supported by credible clinical evidence rather than intuition alone; when the strength of the justification is proportionate to the AI's demonstrated reliability in the relevant circumstances; and when the response is proportionate to the potential consequences of error, delay, and intervention. Where uncertainty can be reduced through additional testing, consultation, or monitoring, those measures may be preferable to immediately choosing between the AI and physician judgment. Physicians also remain accountable for their decisions and should, when circumstances permit, be able to explain and document the patient-specific basis for an override.

Ultimately, a physician's ethical authority to override a more accurate diagnostic AI does not arise simply from professional status or from the fact that the physician is human. It arises when the physician can provide sufficient, patient-specific evidence that the AI's demonstrated general reliability should not control the individual decision. This framework preserves the ethical importance of diagnostic accuracy while recognizing that responsible diagnosis requires more than population-level performance: it requires determining whether that performance applies to the particular patient before whom the decision is being made.

7. References

[1] Hoppe, John Michael et al. "ChatGPT With GPT-4 Outperforms Emergency Department Physicians in Diagnostic Accuracy: Retrospective Analysis." Journal of medical Internet research vol. 26 e56110. 8 Jul. 2024, doi:10.2196/56110

[2] Šimundić, Ana-Maria. "Measures of Diagnostic Accuracy: Basic Definitions." EJIFCC vol. 19,4 203-11. 20 Jan. 2009

[3] Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM; Cochrane Diagnostic Test Accuracy

Working Group. Systematic reviews of diagnostic test accuracy. Ann Intern Med. 2008 Dec 16;149(12):889-97. doi: 10.7326/0003-4819-149-12-200812160-00008. PMID: 19075208; PMCID: PMC2956514.
[4] Saber Tehrani, Ali S et al. "25-Year summary of US malpractice claims for diagnostic errors 1986-2010: an analysis from the National Practitioner Data Bank." BMJ quality & safety vol. 22,8 (2013): 672-80. doi:10.1136/bmjqs-2012-001550
[5] Insani, Widya N et al. "Prevalence of adverse drug reactions in the primary care setting: A systematic review and meta-analysis." PloS one vol. 16,5 e0252161. 26 May. 2021, doi:10.1371/journal.pone.0252161
[6] Mago R, Fagiolini A, Weiller E, Weiss C. Understanding the emotions of patients with inadequate response to antidepressant treatments: results of an international online survey in patients with major depressive disorder. BMC Psychiatry. 2018 Feb 5;18(1):33. doi: 10.1186/s12888-018-1625-y. PMID: 29402325; PMCID: PMC5800100.
[7] Rolfe, Alix et al. "Interventions for improving patients' trust in doctors and groups of doctors." The Cochrane database of systematic reviews vol. 2014,3 CD004134. 4 Mar. 2014, doi:10.1002/14651858.CD004134.pub3
[8] Riley, Richard D et al. "Summarising and validating test accuracy results across multiple studies for use in clinical practice." Statistics in medicine vol. 34,13 (2015): 2081-103. doi:10.1002/sim.6471
[9] Akobeng, Anthony K. "Understanding diagnostic tests 1: sensitivity, specificity and predictive values." Acta paediatrica (Oslo, Norway : 1992) vol. 96,3 (2007): 338-41. doi:10.1111/j.1651-2227.2006.00180.x
[10] Robinson, Eric J et al. "Physician vs. AI-generated messages in urology: evaluation of accuracy, completeness, and preference by patients and physicians." World journal of urology vol. 43,1 48. 27 Dec. 2024, doi:10.1007/s00345-024-05399-y
[11] Khoruzhaya, Anna N et al. "Standalone AI Versus AI-Assisted Radiologists in Emergency ICH Detection: A Prospective, Multicenter Diagnostic Accuracy Study." Journal of clinical medicine vol. 14,16 5700. 12 Aug. 2025, doi:10.3390/jcm14165700
[12] University of Florida College of Medicine. Doc-in-the-Box: Diagnostic Reasoning and Cognitive Biases. University of Florida College of Medicine, 2012, https://clerkship.medicine.ufl.edu/files/2012/07/Doc-in-the-Box-Handout-7-17-12.pdf
[13] Diagnostic Reasoning." ScienceDirect Topics, Elsevier, https://www.sciencedirect.com/topics/medicine-and-dentistry/diagnostic-reasoning
[14] Shikino, Kiyoshi et al. "Evaluation of ChatGPT-Generated Differential Diagnosis for Common Diseases With Atypical Presentation: Descriptive Research." JMIR medical education vol. 10 e58758. 21 Jun. 2024, doi:10.2196/58758
[15] Zwaan, Laura et al. "Relating faults in diagnostic reasoning with diagnostic errors and patient harm." Academic medicine : journal of the Association of American Medical Colleges vol. 87,2 (2012): 149-56. doi:10.1097/ACM.0b013e31823f71e6
[16] Ivanovic, Vladimir et al. "Actionable strategies to minimize diagnostic errors in radiology." The neuroradiology journal vol. 39,4 (2026): 420-427. doi:10.1177/19714009251395719
[17] Berner, Eta S., and Mark L. Graber. "Overconfidence as a Cause of Diagnostic Error in Medicine." The American Journal of Medicine, vol. 121, no. 5, suppl. 1, 2008, pp. S2–S23. https://doi.org/10.1016/j.amjmed.2008.01.001
[18] McLeod, Graham A., et al. "Distinct Visual Biases Affect Humans and Artificial Intelligence in Medical Imaging Diagnoses." npj Digital

Medicine, vol. 9, 2026, article 62. Nature, https://doi.org/10.1038/s41746-025-02226-5

[19] Hall, Jeff. "Study: Generative AI Model Reduces CXR Reading Time by 42 Percent." Diagnostic Imaging, 12 Mar. 2025, www.diagnosticimaging.com/view/study-generative-ai-model-reduces-cxr-reading-time-by-42-percent

[20] Temmen, Sander E et al. "Duration and accuracy of automated stroke CT workflow with AI-supported intracranial large vessel occlusion detection." Scientific reports vol. 13,1 12551. 2 Aug. 2023, doi:10.1038/s41598-023-39831-x

[21] Do, Kyung-Hyun et al. "The Growing Problem of Radiologist Shortages: Korean Perspective." Korean journal of radiology vol. 24,12 (2023): 1173-1175. doi:10.3348/kjr.2023.1010

[22] Huang, Fuxiang, et al. "A Versatile Foundation Model for AI-enabled Mammogram Interpretation." arXiv, 2025, arXiv:2509.20271. https://doi.org/10.48550/arXiv.2509.20271

[23] Chishti, Shahrukh et al. "Artificial Intelligence-Based Differential Diagnosis: Development and Validation of a Probabilistic Model to Address Lack of Large-Scale Clinical Datasets." Journal of medical Internet research vol. 22,4 e17550. 28 Apr. 2020, doi:10.2196/17550

[24] Hager, Paul et al. "Evaluation and mitigation of the limitations of large language models in clinical decision-making." Nature medicine vol. 30,9 (2024): 2613-2622. doi:10.1038/s41591-024-03097-1

[25] Varkey, Basil. "Principles of Clinical Ethics and Their Application to Practice." Medical Principles and Practice, vol. 30, no. 1, 2021, pp. 17–28. Karger, https://doi.org/10.1159/000509119

On The Impossibility of a Neutral Optimizer

Konstin Lee[1]

[1]The Stony Brook School, New York, United States

Abstract

In voting theory and machine learning, impossibility theorems are often taken to show that some unfairness is unavoidable and therefore, to an extent, excusable. However, we argue that this interpretation reverses the logic as an impossibility theorem is a conditional whose antecedent specifies the conditions under which a set of fairness criteria cannot all be simultaneously satisfied. Those conditions also tell us what would need to change to resolve the conflict.

We distinguish three kinds of impossibility claim which include logical, contingent and formalization gaps. Results in algorithmic fairness are generally contingent, yet are discussed as though they were logical. Following Sen's reading of Arrow, we argue that these results concern the information a formalism permits a fairness rule to use, rather than fairness itself. More information can resolve particular conflicts, but every such expansion requires a value judgment that the added information cannot supply. We refer to this as the conservation of normativity.

We then identify five points at which value judgments enter an optimization and argue that fairness audits rarely inspect the fifth, the search procedure. We define a neutral optimizer as a search procedure that adds no value judgment of its own, leaving every value judgment in the problem specification. Neutrality requires three conditions, each of which fails under specified circumstances. Electoral districting is an example in which all three kinds appear together and in which the Supreme Court in *Rucho v. Common Cause* adopted the license reading in declining jurisdiction.

Keywords. Algorithmic fairness, impossibility theorems, social choice, optimization, gerrymandering, procedural legitimacy

1 Introduction

In 2019, the Supreme Court held that claims of partisan gerrymandering raise political questions beyond the scope of the federal courts [1]. However, the Court did not suggest that partisan gerrymandering is acceptable. Indeed, it acknowledged that the challenged maps conflicted with democratic principles. Its concern, instead, was that no judicially manageable standard existed for deciding when partisan advantage becomes excessive. Any such standard requires a baseline of fair representation, yet several plausible baselines exist, and the Constitution does not explicitly select any. The Court thus concluded that choosing among these competing conceptions is a political challenge rather than a judicial one.

Setting aside whether the decision was legally correct, the Court conceded that fairness allows several defensible formalizations, and that none of them is uniquely right. It then treated that concession as a reason not to act. We call this the license reading of impossibility, in which, because the criteria cannot all be satisfied simultaneously, no particular failure to satisfy them is anyone's fault.

The license reading extends beyond constitutional law. As Chouldechova showed, a classifier cannot equalize positive predictive value, false positive rates, and false negative rates across two groups at once when those groups have different base rates [2]. Kleinberg, Mullainathan and Raghavan showed that a risk score cannot be calibrated within groups and balanced across groups for both outcome classes at once, unless prediction is perfect or the base rates are equal [3]. Although both results are rigorous theorems, they are commonly summarized as the claim that achieving fairness on all counts is mathematically impossible, which encourages three further claims. One must simply choose a definition, any reasoned choice is defensible, and the resulting disparities are the price of mathematics.

The summary is misleading because omitting the antecedent turns a claim about a particular state of the world into a claim concerning the limitations of fairness itself. Restoring the antecedent makes the theorem a list of the conditions that produce the conflict, and thus a guide to what would have to change for it to disappear.

We make three claims. First, fairness impossibilities fall into three kinds distinguished by the assumptions they are based on. The best-known algorithmic results belong to the weakest kind, yet they are discussed as though they have the strongest force. Second, escaping those results does not remove the underlying value judgment. It moves the judgment elsewhere. Third, the judgment tends to move to one particular place. Value judgments enter an optimization at five distinct points, or loci, and fairness audits rarely examines the fifth, which is the search procedure. A search procedure that adds no value judgment of its own is a neutral optimizer. Section 2 defines that idea as three conditions and argues that the procedures in general use satisfy none of them.

Section 1.1 sets out the three kinds, and Section 1.4 reads the two canonical results back onto their antecedents. Section 2 locates the five loci and shows why neutrality fails. Section 3 considers electoral districting as a case where all three kinds appear together. Section 4 asks what a deployed optimization owes the people subject to it.

The optimization facts in Section 2 are standard, and our contribution is an interpretation of them rather than a verdict against formal methods. Green traces impossibility results to the narrowness of formal fairness and argues for a substantive alternative [4]. We agree with much of that account but differ in three respects. Our kinds distinguish claims within formal fairness rather than opposing formalism as such. Our diagnosis identifies the additional information that resolves a given conflict, whereas widening the frame establishes only that the frame is narrow. Finally, our argument proceeds within the formalism, because it is within that formalism that institutions act.

1.1 Three impossibilities

Every impossibility theorem comes with conditions that fix when the conflict holds. The three kinds differ in what those conditions are made of.

1.1.1 Logical impossibility

The conditions are formal features of the problem's setup. Here the hypotheses, rather than the conclusions, do the relevant work, and each hypothesis concerns the setup rather than the world. Arrow's theorem needs at least three alternatives and a transitive social ordering over an unrestricted domain [5]. Gibbard–Satterthwaite needs a range of at least three alternatives [6]. Arrhenius's population ethics theorems need the betterness ordering to be complete and transitive [7]. Resolving a conflict of this kind requires changing the problem or abandoning one of its requirements. Changing the world cannot resolve it because the world does not appear in the conditions.

1.1.2 Contingent impossibility

The conflict remains an exact theorem, but its conditions are empirical facts about a particular population, true in one setting and possibly false in another. Chouldechova's result applies when prevalence differs between groups. The result of Kleinberg, Mullainathan and Raghavan applies when base rates differ and prediction is imperfect. Chen and Rodden's districting conclusion holds when a party's supporters are unevenly spread across space [8]. Resolving a conflict of this kind requires changing the world or changing what is measured. The theorem remains true, but no longer applies.

1.1.3 Formalization gaps

Here there are no conditions to change because the difficulty is not a clash between formal requirements. It is the distance between a moral idea and its formal representation. The individual fairness of Dwork et al. [9] needs a task-specific similarity metric, which the framework assumes rather than provides. Ideas like risk, creditworthiness, and teacher effectiveness cannot be observed directly, and the quantities standing in for them can fail in ways no fairness metric detects [10]. No change in the world closes a formalization gap, because no theorem has yet been specified.

Although a formalization gap is not itself an impossibility result, it belongs in this taxonomy as the three kinds classify diagnoses rather than theorems. They capture what may be meant by the claim that fairness is impossible in a given setting. Two diagnoses cite a theorem and differ in the nature of its conditions. The third identifies the absence of a theorem because the moral idea has not been specified precisely enough to yield one. Frequently confused with the other two, this third diagnosis is explored in Section 3, where a court's error in this regard leads to jurisdictional implications.

1.2 Sen's reading of Arrow

Arrow's theorem is commonly read as showing that democratic aggregation is incoherent. Sen has argued at length that this is the wrong lesson [11].

On Sen's account the impossibility comes from the information the framework lets the rule use. That information is thin, consisting of rankings only, with no way to compare how much different people care. Independence of irrelevant alternatives is what makes it thin. It says that society's ranking of a pair may depend only on how individuals rank that same pair. So it rules out how strongly anyone feels, and it rules out any comparison of welfare between persons.

The restriction becomes apparent as a voter mildly prefers x to y, while another would be ruined by y, but under independence the rule sees the same thing in both cases, a preference for x over y, and must treat them alike. The distinguishing information is present in the world and could be elicited, although the framework declares it inadmissible.

Thus, the theorem concerns the limits of a ranking-only picture, which contains too little information to support a complete, transitive, non-dictatorial social ordering, rather than the incoherence of collective choice itself.

The constructive half of Sen's argument reinforces this diagnosis, because possibility results emerge when the rule may compare welfare across people. Different assumptions about such comparisons yield different social orderings. Cardinal full comparability aligns with utilitarian aggregation, while ordinal level comparability aligns with leximin, so the remaining choice between them is an ethical one.

1.3 Information a fairness rule may use

Algorithmic fairness also has the same structure. Take a binary outcome Y, a protected attribute A, a score S, and a prediction $\hat{Y}$. On the framework of Barocas, Hardt and Narayanan [12], almost every fairness criterion in the literature is one of three conditions on how these four variables are jointly distributed, or a weakening of one. The conditions are independence ($\hat{Y} \perp A$), separation ($\hat{Y} \perp A \mid Y$), and sufficiency ($Y \perp A \mid S$). Demographic parity is independence. Equalized odds is separation. Group calibration is sufficiency.

The rule may use those four variables and nothing else. There is no record of what causes Y, no record of the mechanism connecting A to Y, nothing about welfare, and nothing measured over time. The impossibility theorems are theorems about those four variables. They report the limits of a four-variable picture.

1.4 Reading conditions

Proposition 1 (Chouldechova [2])**.** *Let p be the prevalence of the positive class within a group, and let* PPV*,* FPR *and* FNR *be that group's positive predictive value, false positive*

rate and false negative rate. Then

$$\text{FPR} \;=\; \frac{p}{1-p} \cdot \frac{1-\text{PPV}}{\text{PPV}} \cdot (1-\text{FNR}). \tag{1}$$

Proof. In a group of size N there are pN actual positives, so $\text{TP} = pN(1-\text{FNR})$. From $\text{PPV} = \text{TP}/(\text{TP}+\text{FP})$ we get $\text{FP} = \text{TP}(1-\text{PPV})/\text{PPV}$. Dividing by the $(1-p)N$ actual negatives gives (1). □

Suppose two groups have the same predictive value and the same false negative rate, but different prevalences. Then the right-hand side of (1) differs between them, so their false positive rates must differ too. That is the impossibility.

The license reading obscures the difference between mathematical necessities and features of American policing. The ratio $p/(1-p)$ is prevalence, it is measured, and the algebra does not determine its value.

Theorem 2 (Kleinberg, Mullainathan and Raghavan [3])**.** *Consider a risk score on two groups. Calibration within groups says that among those assigned score v, a v fraction are positive. Balance for the negative class says the average score among negatives is the same in both groups. Balance for the positive class says the same for positives. These three hold at once only if prediction is perfect, or the two groups have equal base rates.*

We fix a group with base rate p. Calibration says $\mathbb{E}[Y \mid S = v] = v$, so the group's average score equals its base rate, that is $\mathbb{E}[S] = \mathbb{E}[\mathbb{E}[Y \mid S]] = \mathbb{E}[Y] = p$. Split that average by outcome class to get $\mathbb{E}[S] = p\,\mathbb{E}[S \mid Y{=}1] + (1-p)\,\mathbb{E}[S \mid Y{=}0]$. Write $x_1 = \mathbb{E}[S \mid Y{=}1]$ and $x_0 = \mathbb{E}[S \mid Y{=}0]$. The two balance conditions say these are the same in both groups. So each group satisfies $p = px_1 + (1-p)x_0$, that is

$$\frac{p}{1-p} \;=\; \frac{x_0}{1-x_1}.$$

The right side is the same for both groups. So p must be too, unless the expression degenerates, which happens exactly when $x_0 = 0$ and $x_1 = 1$. That is perfect prediction. Those two escapes are the only two ways the equation can fail to bind. The authors also prove an approximate version, which closes off the retreat to being nearly fair on all three counts.

Both results are therefore of the same kind. The conditions are unequal base rates and imperfect prediction, neither of which is a law of nature.

1.5 Escapes and the conservation of normativity

When an impossibility arises from the rule's limited information, one response is to let it use more information.

The most demanding version of this adds causal structure since counterfactual criteria ask that a prediction stay the same if we imagine changing a person's protected attribute.

Path-specific criteria go further and sort the variables in between into two classes, those that may legitimately carry influence from the protected attribute to the outcome, and those that may not. The observational impossibility theorems do not apply to these criteria at all. But writing one down means supplying a causal graph, and sorting the variables means deciding which pathways count as legitimate. Those are ethical judgments encoded as modelling assumptions, and they are contested. Kohler-Hausmann argues that treating race as a manipulable cause misunderstands what race is, since on a constitutive account it is not an attribute you can vary while holding everything else fixed [13].

A second version widens the range of groups that a criterion has to cover. Multicalibration asks for calibration across every subgroup in a rich, computationally identifiable family at once. It addresses fairness gerrymandering, where a classifier satisfies a criterion for each protected attribute on its own and fails where those attributes intersect. These are genuine possibility results. They also move the question rather than answer it, because someone now has to choose the family of subgroups, and no formalism makes that choice.

Call an escape informational when it resolves a conflict by letting the rule see more. Every informational escape demands a judgment that the new information does not settle. So the value judgment is carried into the new information rather than discharged by it. Letting Arrow's rule compare welfare leaves you choosing between utilitarian and leximin aggregation. Adding causal structure leaves you specifying a graph and nominating legitimate pathways. Widening the family of subgroups leaves you choosing the family, which we call the conservation of normativity.

This claim concerns informational escapes because abandoning a requirement is different from relocating a judgment. Abandoning transitivity removes a commitment instead of transferring it, hence the result is a shorter list of requirements, not a displaced judgment. More generally, no escape is costless. Informational escapes relocate the judgment, while escapes by abandonment must justify the abandonment.

The same point appears in a result with the opposite formal structure. Lan and Chiang give five axioms for measuring fairness in resource allocation and derive a uniqueness theorem, under which one single-parameter family of measures satisfies all five [14]. Where Arrow shows that plausible axioms allow no aggregator at all, they show that plausible axioms allow exactly one family, indexed by a free parameter β that sets how much inequality the measure tolerates. The axioms do not fix β. Impossibility and uniqueness point in formally opposite directions and leave the same residue. Formal analysis can narrow the space of defensible fairness rules. It cannot choose within it.

2 Entrance of value judgments

On the standard account, ethics lives in the objective function. Optimization methods are powerful and indifferent, in that they find a good solution to whatever problem you pose, and the values are all in the posing. That is the account we have assumed so far. It is incomplete.

Call a search procedure a neutral optimizer when it adds no value judgment of its own. On that picture every value judgment in a system lives in how the problem was specified, and the search is a mere instrument, indifferent between anything the specification does not distinguish. Neutrality needs three conditions to hold.

N1, reachability. Every compromise the specification leaves open can actually be reached by some setting of the procedure's parameters. The procedure rules out nothing the objective and constraints allow.

N2, prior independence. The procedure builds in no assumption about what kind of problem it will be handed.

N3, indifference. Where several solutions look the same under the stated objective, it does not matter to anyone which one the procedure returns.

A neutral optimizer would vindicate the standard account. It would collapse the five loci below into the fourth. Each of the three conditions fails for the procedures in general use, and each failure bears on outcomes for identifiable people.

2.1 Five loci

2.1.1 Framing

What counts as a decision, where the boundary of the system lies, and whether the decision should be made at all. This is the territory mapped by Selbst et al. [15], whose framing and solutionism traps name the failure modes. Fairness properties also fail to compose, in that parts which each satisfy a criterion can combine into a whole that violates it.

2.1.2 Population and unit of analysis

Who is in the reference group, over what stretch of time, and at what level of aggregation. A disparity visible at one level of aggregation can vanish at another, so this choice fixes what an audit is capable of finding.

2.1.3 The measured

The step from an idea you cannot observe, such as risk or creditworthiness, to a number you can compute. Jacobs and Wallach supply the apparatus [10]. Construct validity has several components, one of which is consequential validity, where measurement theory itself says that whether a measure is adequate is partly an ethical question.

2.1.4 Objective, constraints, and weights

The one place the standard account already looks. Everything that account says about values entering a system is said here.

2.1.5 Search procedure

Which solutions can be reached, and which of several equally good ones comes back.

In fairness audits, the fifth locus is seldom examined. Such audits typically assess the training data, outcomes, error rates for each group, and occasionally the objective. However, how the model was found is rarely documented. Champion has argued that the choice of optimization strategy in fairness problems is itself value laden [16]. We agree with this perspective, and the subsequent sections will articulate this argument in terms of what a procedure can and cannot return. The remainder this section will be spent addressing N1, N2 and N3 sequentially.

2.2 Reachability

Take m objectives $f_1, \ldots, f_m$ to be minimized over a feasible set, and let $\mathcal{Z} \subseteq \mathbb{R}^m$ be the set of achievable objective vectors. A point $z \in \mathcal{Z}$ is Pareto optimal if nothing else in $\mathcal{Z}$ is at least as good on every objective and strictly better on one. The Pareto front is the set of such points. Not every point on the front is ethically defensible, as maximum accuracy with no attention to fairness can sit on it. However, every defensible compromise sits on it, because anything off it loses on some objective without gaining on another. Call z supported if it minimizes $\sum_i \lambda_i z_i$ over $\mathcal{Z}$ for some weight vector λ with every $\lambda_i > 0$, and unsupported otherwise. The usual way to explore a trade-off is to minimize such a weighted sum and vary the weights.

Proposition 3. (i) *Every minimizer of $\sum_i \lambda_i z_i$ over $\mathcal{Z}$ with all $\lambda_i > 0$ is Pareto optimal.* (ii) *If $\mathcal{Z} + \mathbb{R}^m_+$ is convex, every Pareto optimal point is supported.* (iii) *Without convexity,* (ii) *can fail. There are non-convex sets $\mathcal{Z}$ containing Pareto optimal points that are unsupported.*

Proof. For (i), suppose $z^\star$ minimizes the weighted sum and some $z' \in \mathcal{Z}$ dominates it. Then $\sum_i \lambda_i z'_i < \sum_i \lambda_i z^\star_i$, since every $\lambda_i > 0$ and at least one coordinate strictly decreases, contradicting minimality. Part (ii) is the standard supporting hyperplane argument [17]. For (iii) take $\mathcal{Z} = \{(0,3), (2,2), (3,0)\}$. The point $(2,2)$ is Pareto optimal, since neither other point dominates it. For $(2,2)$ to minimize $\lambda_1 z_1 + \lambda_2 z_2$ we would need $2\lambda_1 \leq \lambda_2$ and $2\lambda_2 \leq \lambda_1$, which together give $4\lambda_1 \leq \lambda_1$, impossible for $\lambda_1 > 0$. So $(2,2)$ is unsupported. □

Non-convexity alone does not guarantee the presence of unsupported points. However, it eliminates the guarantee that all Pareto points are supported. For a finite feasible set, $\mathcal{Z}$ is a finite collection of points, each contained in its own convex hull. Therefore, we ask whether a Pareto point lies on the lower-left boundary. In the example above, $(2,2)$ lies strictly above the line joining $(0,3)$ and $(3,0)$.

An exception must be acknowledged as it is applicable in some contexts but not in others. If the practitioner employs a lottery over models instead of a single model, the achievable set becomes the convex hull of $\mathcal{Z}$ and part (ii) is applicable once more. A fair

coin between $(0, 3)$ and $(3, 0)$ gives expected losses $(1.5, 1.5)$, which beats $(2, 2)$ on both objectives, so once mixtures are allowed $(2, 2)$ is not Pareto optimal at all. This scenario is general as opposed to being specific to this example, since an unsupported point is one that lies above the lower-left boundary. Hence the failure of N1 is conditional on the deployed object being a single rule rather than a lottery over rules.

The condition holds in the contexts discussed in this paper. A districting authority enacts one map, and a court reviews the map that was enacted. In cases where individual decisions are at stake, resolving them by a draw is contested, since it bases outcomes on factors unrelated to the specific case case. We do not assert that objection is decisive. Instead, we claim that reachability is achieved through randomization instead of being inherently available, aligning with the predictions of the conservation of normativity.

This directly impacts the accuracy–fairness trade-off. Generally, one minimizes $\mathcal{L}_{\text{acc}} + \lambda \mathcal{L}_{\text{fair}}$ and adjusts λ. Discrete model classes provide no reason for $\mathcal{Z}$ to be convex. However this is an absence of reason rather than a demonstration, and whether unsupported points exist in a given problem is a question about that problem which can be settled by computing the front directly. Where they do exist, defensible compromises remain undetected by method, regardless of how carefully the weights are searched. Wei and Niethammer report the same limitation as a matter of method, observing that many existing fairness algorithms perform linear scalarization and that this restricts which Pareto optimal solutions can be recovered [18]. The practitioner's entire discussion of λ then takes place over a set the method has quietly restricted. The restriction goes unreported, since what was left off was never computed.

The weighted Chebyshev scalarization reaches points where a weighted sum cannot. It minimizes

$$\max_i \ \omega_i (f_i(x) - z_i^\star)$$

over feasible x, where the $\omega_i > 0$ are weights and $z^\star \in \mathbb{R}^m$ is a reference point no worse in every coordinate than anything achievable. Every Pareto optimal point, supported or not, minimizes this for some ω. The plain version can return weakly Pareto optimal points, which is why a small extra term $\rho \sum_i (f_i(x) - z_i^\star)$ is added [17]. These functions are standard in multiobjective optimization and still uncommon in fairness research.

2.3 No assumption-free search

Average over all objective functions on a finite search space, weighting each equally, and every non-repeating search algorithm performs the same as every other [19].

That uniform average does not model the problems practitioners face. Instead, the theorem shows that if an algorithm outperforms random search for a class of problems, its advantage derives from an assumption about that class built into its structure,[1] and it is paid for by worse performance elsewhere. No search procedure is assumption-free.

[1] Such as good solutions clustering together, the objective being smooth, or gradients being informative.

2.4 Multiplicity

Breiman named the Rashomon effect [20], in which a dataset regularly admits many models with nearly identical accuracy and quite different internal structure. Marx, Calmon and Ustun made the consequences precise [21]. Fix a tolerance ε and consider every model within ε of the best empirical risk. Ambiguity is the fraction of individuals who get conflicting predictions across that set. They give integer programming methods to compute it exactly for linear classifiers, and find it is substantial on real data, including prediction of recidivism.

Therefore, aggregate performance does not determine who gets what. Among models that are indistinguishable on every metric reported, many identifiable people are treated differently. Which model gets deployed turns on details of the search that fairness audits do not record, namely the initialization, the ordering of the features, and the stopping criterion.

Arbitrariness in a single decision is not necessarily an injustice. Creel and Hellman locate the distinctive wrong in the systematic application of the same rule [22]. When one model, or several closely related models trained on similar data for similar ends, is used across an entire domain, an arbitrary feature counting against a person is applied by every decision-maker in that domain. What might otherwise be a minor setback becomes exclusion from an entire sphere of opportunity. The wrong need not track a protected attribute, and fairness metrics may therefore fail to register it.

2.5 Lack of a neutral optimizer

N1 fails when the achievable set is non-convex, contains unsupported Pareto points, and the practitioner explores the trade-off by varying weights on a scalarized sum, which is standard practice, and on the deployed object being a single rule rather than a lottery over rules. In such cases, defensible compromises are excluded by the method rather than by the specification, and the exclusion goes unreported. N2 fails for any procedure that outperforms random search on the problems it is actually given, which covers the procedures in general use. N3 fails whenever the Rashomon set has non-zero ambiguity, since the procedure then decides how identifiable people are treated on grounds the objective never mentions.

A procedure's selections are not all value judgments. An algorithm has an inductive bias when its structure favours some hypotheses over others, it restricts reachability when solutions permitted by the specification lie beyond its range, and it selects arbitrarily when it returns one of several solutions the objective cannot distinguish. A selection becomes normative when the alternatives differ in their treatment of identifiable people and the grounds of selection are absent from the specification. N1 and N3 meet that test directly, unsupported Pareto points are ethically defensible compromises, and Rashomon ambiguity is measured over individuals. N2 does so through N1 and N3, when an inductive bias affects which compromises are reachable or which of several equally good models is returned.

Formal analysis can narrow the space of defensible fairness rules without selecting among them. Thus, the final decision must be made where an institution can be held accountable. Section 1.5 reaches the same conclusion from impossibility and uniqueness results.

This is a contingent, rather than logical, impossibility. Nothing above implies that a procedure satisfying N1–N3 would be incoherent, and some problems admit one. Each condition also identifies its own escape. Chebyshev scalarization with the extra term restores N1. No procedure restores N2, though its hidden assumption can be stated, and reporting model-class ambiguity converts an N3 failure from a silent choice into a disclosed one.

3 Districting

Districting brings all three kinds together and makes the fifth locus visible inside a courtroom. Section 3.2 applies the three kinds to a piece of judicial reasoning.

3.1 The fifth locus in the courtroom

Modern outlier analysis samples rather than selects a single fairness metric. The analyst generates many districting plans that satisfy the state's stated criteria, derives the distribution of partisan outcomes across that collection, and asks whether the enacted plan is extreme relative to it. This approach is practical because of the recombination family of Markov chains introduced by DeFord, Duchin and Solomon [23], which merges two neighboring districts, draws a random spanning tree of the combined unit, and cuts one edge to produce two population-balanced pieces. Its predecessor, the boundary flip chain, moves individual precincts across district lines. It mixes far more slowly and concentrates on plans with unrealistically long boundaries, which forces the analyst to add compactness weights of arbitrary size.

The two chains explore different regions of the space of plans, producing collections with different distributions of compactness and partisan outcome. A map that is a mild outlier under one can be extreme under the other, and no canonical neutral collection is available. Sampling uniformly from the space of balanced connected partitions is computationally infeasible, the practical chains have no proven mixing times at realistic scales, and rival frameworks are actively defended.

Choosing between chains on computational-efficiency grounds changes the reference distribution against which the enacted map is judged and can therefore change a conclusion about its lawfulness. Each constraint placed on the collection has the same character.[2] The neutral baseline encodes a districting philosophy, and it is selected by an expert witness.

3.2 Decision the theorems do not make

The consolidated appeals in *Rucho* involved a Republican gerrymander in North Carolina and a Democratic one in Maryland. The decision is therefore not readily characterized as

[2]Population deviation tolerance, the compactness measure and its threshold, and whether county splits are penalized.

partisan, and its reasoning must be assessed on its own terms. The Court found the claims nonjusticiable for want of a manageable standard, and attributed that want to the many defensible conceptions of fair representation.

On our classification, what the Court found was a formalization gap. There is no single formal version of fair representation, the candidate versions build in contestable political judgments, and the Constitution picks none of them. The Court then treated that gap as though it were a logical impossibility, as though something had been proved unachievable, and declined jurisdiction on that basis.

The plaintiff's argument did not necessitate closing the gap. Instead, their evidence was statistical in nature, demonstrating that the challenged plan was exceptionally partisan compared with alternative maps . The enacted map was compared against a large collection generated from the state's own criteria, and the court was asked whether it significantly deviated from the resulting distribution. That questions the extremity relative to a state's declared standards. It does not require deciding which conception of fair representation is correct, much as spotting an outlier does not require knowing the true mean. Justice Kagan's dissent presses the same point, noting that lower courts had been applying a three-part inquiry into intent, effect, and justification without evident difficulty.

This analysis does not resolve whether *Rucho* was rightly decided, but its reasoning illustrates the equivocation at issue, in which a formalization gap became a jurisdictional barrier. Other courts treated the gap as tractable. The Pennsylvania Supreme Court invalidated that state's congressional map under its Free and Equal Elections Clause and imposed a manageable standard where the federal court had found none. The *Rucho* Court itself indicated state constitutions as viable alternative route.

3.3 Unstated escape condition

The formal literature on districting fairness assumes single-member districts with plurality winners, including the impossibility results, the findings about geographic bias, the competing metrics such as the efficiency gap [24], and the ensemble methods. That assumption is what gives the map its power in the first place. Under multi-member districts with proportional or single transferable voting, seats track votes far more closely, and the leverage available from drawing lines falls sharply. The tension between compactness and proportionality arises from the electoral system rather than districting itself. The United States used multi-member congressional districts at various points before a statute prohibited them, so this alternative is historically concrete rather than merely theoretical.

The license reading treats partisan advantage in map drawing as regrettable but mathematically inevitable, leaving a choice of metric and its consequences. A reading that foregrounds escape conditions instead identifies the single-member plurality district as the source of the situation and notes that Congress could repeal the statute that created it. Both readings rest on the same theorems, but only the latter identifies an intervention beyond further map optimization.

4 Optimization to legitimacy

Lan and Chiang derive a unique family of fairness measures and stop just short of the question of who applies the measure, and whether the people affected have any say [14].

When no arrangement is fair under every reasonable interpretation, the central question becomes not whether an outcome is fair, but what gives its resolution authority rather than mere force, which is a question of legitimacy. Although participation matters, it alone is insufficient. Sloane et al. show how participatory methods can become a design patch, supplying the appearance of consent without transferring decision-making power [25]. Following Estlund, legitimacy instead belongs to a procedure that can be justified to the people bound by it in terms they can accept [26], and that standard does not require a uniquely correct answer, which the preceding sections argue is unavailable.

4.1 Requirements of the standard

Applied to a deployed optimization, that standard yields five procedural requirements that do not claim to settle the underlying ethical question.

The first requirement is to state the trade-off by identifying the conflicting criteria and the local conditions under which they conflict, so that readers can distinguish logical from contingent conflict. The second is to defend the choice, because a point on the trade-off, whether set by β, λ, a threshold, or a family of subgroups, must be presented and defended as a choice rather than passed off as the method's output. The third is to disclose multiplicity, so where near-optimal alternatives disagree about identifiable people, model-class ambiguity should be reported alongside aggregate performance. The fourth is to report the escape conditions, since the conditions that make the conflict bind also identify interventions that would resolve it, often outside the audited system. The fifth is to permit challenge and revision, because people subject to the decision need a route to contest the second choice, and that choice must be revisable without rebuilding the entire system.

Requirements 1 and 4 come out of reading theorems as conditionals, and requirement 3 out of Section 2.4. Requirements 2 and 5 are where the conservation of normativity gets discharged institutionally, since the judgment it moves has to end up somewhere it can be argued with. Requirements 2 and 3 also happen to be the disclosures that a failure of neutrality makes necessary. Where N1 fails, which scalarization was used belongs under requirement 2, right next to the choice of weights, because together they fix which compromises were ever available. Where N3 fails, the ambiguity of the model class belongs under requirement 3, because aggregate accuracy hides it. Neither disclosure makes the optimizer neutral. Both make its non-neutrality answerable.

These procedural requirements leave the substantive choice of β or λ unresolved, but Lan and Chiang offer partial guidance. Rather than asking people to state a fairness parameter, they elicit a short series of binary comparisons between concrete allocations, for example whether $[50, 50]$ is less fair than $[58, 42]$, and infer the compatible parameter values. Although a few comparisons can narrow the range considerably, preferences over abstract

number pairs are not consent to actual allocations, and aggregating elicited parameters across people remains a social-choice problem. Even so, the remaining parameter concerns judgments people hold and can articulate without technical expertise, which makes requirement 2 actually operational.

5 Conclusion

With its antecedent restored, a fairness impossibility theorem states the conditions that produce a conflict.[3] It also identifies what would have to change for that conflict to resolve.

The value judgment does not disappear, as letting the rule use more information relocates rather than settles the contested judgment, and relocation can make it less visible. A judgment moved from an unstated assumption about base rates to an explicit claim about legitimate causal pathways becomes open to argument. One moved into a Markov chain selected for its mixing time may become obscured, while continuing to shape the treatment of identifiable people. Section 3.1 illustrates this dynamic in court. The requirements in Section 4.1 therefore concern transparency and accountability. The remaining choice should be visible, made by someone answerable for it, and open to revision by the people it binds.

The title follows the same principle as there is no neutral optimizer, not because neutrality is incoherent, but because its three conditions fail under specified circumstances, two of which can be addressed. A condition-free impossibility claim would reproduce the very error this paper identifies.

In *Rucho* a court found that fairness in districting allows several defensible formalizations and concluded that there was nothing they could do. In Pennsylvania a court reached the same finding and proceeded anyway. They were applying different authorities, the federal political question doctrine and a state Free and Equal Elections Clause, and that difference does part of the work, not all of it. *Rucho* did not hold that a plurality of fairness conceptions is manageable under a state clause and unmanageable under federal law. It reasoned from the plurality itself to the absence of any judicially manageable standard. That step is not a legal one, and no theorem cited here supports it.

References

[1] *Rucho v. Common Cause*, 588 U.S. ___, 139 S. Ct. 2484 (2019).

[2] A. Chouldechova, "Fair prediction with disparate impact: a study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.

[3] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *Proc. 8th Innovations in Theoretical Computer Science Conf. (ITCS)*, 2017.

[3] Unequal base rates, imperfect prediction, asymmetric residential clustering, single-member plurality districts.

[4] B. Green, "Escaping the impossibility of fairness: from formal to substantive algorithmic fairness," *Philosophy & Technology*, vol. 35, art. 90, 2022.

[5] K. J. Arrow, *Social Choice and Individual Values*, 2nd ed. New Haven, CT, USA. Yale Univ. Press, 1963. First edition 1951.

[6] A. Gibbard, "Manipulation of voting schemes: a general result," *Econometrica*, vol. 41, no. 4, pp. 587–601, 1973.

[7] G. Arrhenius, "An impossibility theorem for welfarist axiologies," *Economics and Philosophy*, vol. 16, no. 2, pp. 247–266, 2000.

[8] J. Chen and J. Rodden, "Unintentional gerrymandering: political geography and electoral bias in legislatures," *Quarterly Journal of Political Science*, vol. 8, no. 3, pp. 239–269, 2013.

[9] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proc. 3rd Innovations in Theoretical Computer Science Conf. (ITCS)*, 2012, pp. 214–226.

[10] A. Z. Jacobs and H. Wallach, "Measurement and fairness," in *Proc. ACM Conf. on Fairness, Accountability, and Transparency (FAccT)*, 2021, pp. 375–385.

[11] A. Sen, "The possibility of social choice," *American Economic Review*, vol. 89, no. 3, pp. 349–378, 1999.

[12] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. Cambridge, MA, USA. MIT Press, 2023.

[13] I. Kohler-Hausmann, "Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination," *Northwestern University Law Review*, vol. 113, p. 1163, 2018.

[14] T. Lan and M. Chiang, "An axiomatic theory of fairness in resource allocation," manuscript, 2011.

[15] A. D. Selbst, d. boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in *Proc. Conf. on Fairness, Accountability, and Transparency (FAT*)*, 2019, pp. 59–68.

[16] H. Champion, "On values in fairness optimization with machine learning," *Philosophy of Science*, vol. 92, pp. 1372–1382, 2025.

[17] K. Miettinen, *Nonlinear Multiobjective Optimization*. Boston, MA, USA. Kluwer, 1999.

[18] S. Wei and M. Niethammer, "The fairness-accuracy Pareto front," *Statistical Analysis and Data Mining*, vol. 15, no. 3, pp. 287–302, 2022.

[19] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Trans. Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, 1997.

[20] L. Breiman, "Statistical modeling: the two cultures," *Statistical Science*, vol. 16, no. 3, pp. 199–231, 2001.

[21] C. T. Marx, F. P. Calmon, and B. Ustun, "Predictive multiplicity in classification," in *Proc. 37th Int. Conf. on Machine Learning (ICML)*, 2020, pp. 6765–6774.

[22] K. Creel and D. Hellman, "The algorithmic leviathan: arbitrariness, fairness, and opportunity in algorithmic decision-making systems," *Canadian Journal of Philosophy*, vol. 52, no. 1, pp. 26–43, 2022.

[23] D. DeFord, M. Duchin, and J. Solomon, "Recombination: a family of Markov chains for redistricting," *Harvard Data Science Review*, vol. 3, no. 1, 2021.

[24] N. O. Stephanopoulos and E. M. McGhee, "Partisan gerrymandering and the efficiency gap," *University of Chicago Law Review*, vol. 82, no. 2, pp. 831–900, 2015.

[25] M. Sloane, E. Moss, O. Awomolo, and L. Forlano, "Participation is not a design fix for machine learning," in *Proc. 2nd ACM Conf. on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*, 2022, pp. 1–6.

[26] D. Estlund, *Democratic Authority: A Philosophical Framework*. Princeton, NJ, USA. Princeton Univ. Press, 2008.