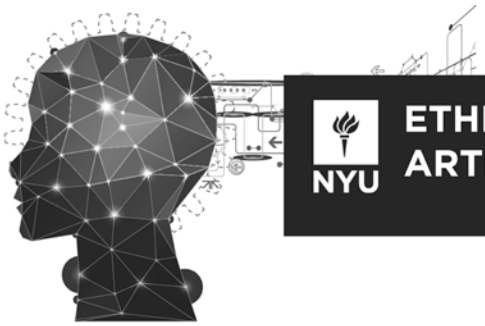


NYU ETHICS OF ARTIFICIAL INTELLIGENCE

FRIDAY OCTOBER 14

Eisner/Lubin Auditorium, 4th Floor, NYU Kimmel Center, 60 Washington Square South

Overflow Rooms (with live stream): Kimmel 405/406, Kimmel 905/907

8:30 - 9:40 am: REGISTRATION AND COFFEE

9:40 am: CONFERENCE OPENING: Thomas J. Carew, Dean of Faculty of Arts and Science, NYU

10:00 am - 12:00 noon: ETHICS OF AI: GENERAL ISSUES

Chair: David Chalmers

Nick Bostrom (Oxford University, Future of Humanity Institute)
Cooperation, Legitimacy, and Governance in AI Development

Virginia Dignum (Delft University of Technology, Technology, Policy and Management)
On Moral Decisions by Autonomous Systems

Yann LeCun (NYU, Data Science; Facebook)
Should We Fear Future AI Systems?

12:00 - 1:30 pm: LUNCH BREAK

1:30 - 3:30 pm: ETHICS OF SPECIFIC TECHNOLOGIES

Chair: Ned Block

Peter Asaro (New School, Media Studies)
Killer Robots and the Ethics of Autonomous Weapons

Kate Devlin (Goldsmiths College, University of London, Computing)
The Ethics of Artificial Sexuality

Vasant Dhar (NYU, Data Science)
Equity, Safety and Privacy in the Autonomous Vehicle Era

Adam Kolber (Brooklyn Law School)
Code is Not the Law: Blockchain Contracts and Artificial Intelligence

3:30 - 4:00 pm: COFFEE BREAK

4:00 - 6:00 pm: BUILDING MORALITY INTO MACHINES

Chair: Matthew Liao

Stephen Wolfram (Wolfram Research)
How to Tell AIs What to Do (and What to Tell Them)

Francesca Rossi (IBM; University of Padova)
Ethical Embodied Decision Making

Peter Railton (University of Michigan, Philosophy)
Machine Morality: Building or Learning?

SATURDAY OCTOBER 15

NYU Cantor Film Center (36 East 8th St), Room 200

Overflow Room (with live stream): Cantor Room 101

8:30 - 9:00 am: REGISTRATION AND COFFEE

9:00 am - 12:00 noon: ARTIFICIAL INTELLIGENCE AND HUMAN VALUES

Chair: David Chalmers

Stuart Russell (UC Berkeley, Computer Science)
Provably Beneficial AI

Eliezer Yudkowsky (Machine Intelligence Research Institute)
Difficulties of AGI Alignment

Meia Chita-Tegmark (Future of Life Institute) and Max Tegmark (MIT, Future of Life Institute)
What We Should Want: Physics and Psychology Perspectives

Wendell Wallach (Yale, Bioethics)
Moral Machines: From Machine Ethics to Value Alignment

Steve Petersen (Niagara University, Philosophy)
Superintelligence as Superethical

12:00 - 1:30 pm: LUNCH BREAK

1:30 - 3:30 pm: MORAL STATUS OF AI SYSTEMS

Chair: Ned Block

S. Matthew Liao (NYU, Bioethics)
Artificial Intelligence and Moral Status

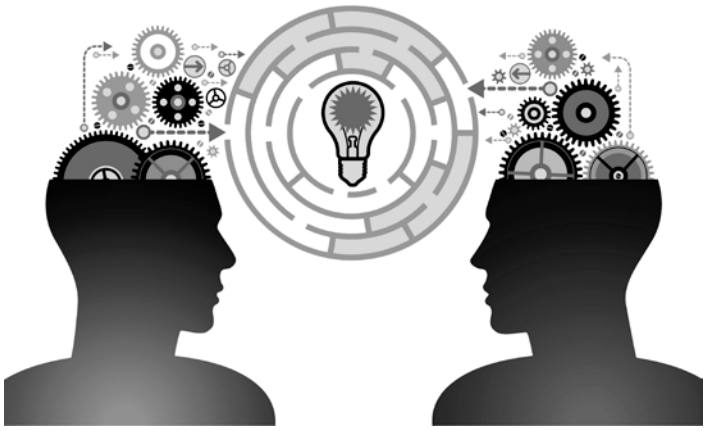
Eric Schwitzgebel (UC Riverside, Philosophy) and Mara Garza (UC Riverside, Philosophy)
The Rights of Artificial Intelligences

John Basl (Northeastern, Philosophy) and Ronald Sandler (Northeastern, Philosophy)
The Case for AI Research Subjects Oversight Committees

3:30 - 4:00 pm: COFFEE BREAK

4:00 - 6:00 pm: PANEL DISCUSSION

Panelists: Daniel Kahneman, Gary Marcus, Susan Schneider, Jürgen Schmidhuber, Jaan Tallinn



NYU | Center for Mind, Brain, & Consciousness

NYU | Center for Bioethics

ETHICS OF AI: GENERAL ISSUES

Nick Bostrom (Oxford University, Future of Humanity Institute)
Cooperation, Legitimacy, and Governance in AI Development

The development of general artificial intelligence would rank among the most important transitions in the history of human civilization, one that may have far-reaching global consequences. Achieving a beneficial outcome will require solving a number of technical and political challenges. The technical challenges have recently started to receive some attention. In this paper, we outline some of the key challenges in the political sphere. We first characterize what an ideal scenario could look like, if practical difficulties are set aside. We then sketch a class of more practically feasible and realistic scenarios that seek to approximate the ideal scenario. We also describe some options available to actors in the different scenarios that could move the development course closer to the ideal scenario.

Virginia Dignum (Delft University of Technology, Technology, Policy and Management)
On Moral Decisions by Autonomous Systems

As robots and other AI systems move from being a tool to being teammates, one of the most important research directions is to rethink the ethical implications of their actions and decisions. Means are needed to integrate moral, societal and legal values with technological developments in Artificial Intelligence, both within the design process as well as part of the deliberation algorithms employed by these systems. In this paper, we propose alternative ways to model ethical reasoning and discuss their consequences to the design of robots and softbots. Depending on the level of autonomy and social awareness of AI systems, different methods for ethical reasoning are needed. Given that ethics are dependent on the socio-cultural context and are often only implicit in deliberation processes, methodologies are needed to elicit the values held by designers and stakeholders, and to make these explicit can lead to better understanding and trust on artificial autonomous systems.

Yann LeCun (Facebook, NYU Data Science)
Should We Fear Future AI Systems?

SPECIFIC TECHNOLOGIES

Peter Asaro (New School, Media Studies)
Killer Robots and the Ethics of Autonomous Weapons

As the militaries of technologically advanced nations seek to apply increasingly sophisticated AI and automation to weapons technologies, a host of ethical questions arise. Central among these is whether it is ethical to delegate the decision to use lethal force to an autonomous system that is not under meaningful human control. Further questions arise as to who. or what, could or should be held responsible when lethal force is used improperly by such systems. I will also explore the efforts of the Campaign to Stop Killer Robots to establish a treaty prohibiting such weapons, how the issue has been discussed and debated within the United Nations, and the role of engineers and scientists in these discussions. The talk will also examine the implications of such a treaty, or the lack of one, for ethical limits on AIs and autonomous systems to make lethal decisions outside of armed military conflict.

Kate Devlin (Goldsmiths College, University of London, Computing)
The Ethics of Artificial Sexuality

One popular application of artificial intelligence is the companion robot - an anthropomorphised cognitive system that can be of use to humans. These exist in basic form but research continues into the goal of a sentient system - one possessing perception, cognition, emotion and action. However, something is missing from current research: sexual behaviour is largely ignored despite its central role in biological and social behaviour. Existing research into sex and robots centres on a superficial exploration of human attachment: a male-dominated approach of machine-as-sex-machine. Sex robots currently take the form of variations on mechanised sex dolls but manufacturers are increasingly moving towards the idea of using AI to create a machine that can learn and alter its behaviour. There are important psychological and social dimensions to sex and sexuality, and they also prompt significant ethical, legal and policy questions. Thinking ethically and politically about how to deal with artificial sexuality forces a wide-ranging reassessment of established legal perspectives on sex. Socially, imagining how to befriend a machine allows a re-examination of how contemporary societies value individuals and groups. Cognitively, how we understand ourselves, one another, and our relationships with the outside world can be reframed in the context of the robot companion. How, ethically, should (or should not) sex robots develop?

Vasant Dhar (NYU Center for Data Science)
Equity, Safety, and Privacy in the Autonomous Vehicle Era

Big data from onboard vehicular systems can be used to help determine liability in accidents, streamline insurance pricing, motivate better driving practices, and improve safety—all with potentially minimal impact on privacy. I discuss the implications of the availability of this new data for accident law and the necessity of policy to limit invasion of privacy.

Adam Kolber (Brooklyn Law School)
Code is Not the Law: Blockchain Contracts and Artificial Intelligence

Earlier this year, the first high-profile decentralized autonomous organization (“TheDAO”) was formed. Running on smart contract and blockchain technology, no central authority owns or controls it. While the technology promises to democratize ownership, a bug in TheDAO was exploited to drain more than $60 million from its value. When a technological solution was proposed to render the exploit ineffective, many denounced the idea arguing that the alleged hacker simply withdrew money in accordance with the organization’s agreed-upon contractual terms (in the form of computer code). Since the contract code is law in their minds, the alleged hacker did nothing wrong.

I argue that, from a legal and ethical perspective, smart contract code should not always be treated as law. If contract code were always law, not-so- smart contracts with bugs could prove particularly dangerous as they interact with an increasingly intelligent Internet of Things. Moreover, while blockchain smart contracts are not very intelligent now, they have high levels of “artificial autonomy” in the limited sense that we cannot do much to rein them in when they go awry. In the more distant future, an A.I. might exercise interpretive judgment to mitigate contracting risks if we are able to load such a system with the right motivational framework.

BUILDING MORALITY INTO MACHINES

Stephen Wolfram (Wolfram Research)
How to Tell AIs What to Do (and What to Tell Them)

Francesca Rossi (IBM, University of Padova)
Ethical Embodied Decision Making

Decision making is an ubiquitous task in our life. However, most of the times humans are not very good at it, because of cognitive biases and data handling difficulties. Intelligent decision support systems are intended to help us in this respect. To build an effective human-machine symbiotic system with the capability to make optimal decisions, we advocate for an embodied environment where humans are immersed. Moreover, we need humans to trust such systems. To build the right level of trust, we need to be sure that they act in a morally acceptable way. Therefore, we need to be able to embed ethical principles (as well as social norms, professional codes, etc.) into these systems. Existing preference modelling and reasoning framework can be a starting point, since they define priorities over actions, just like an ethical theory does. However, it is not yet clear how to mix preferences (that are at the core of decision making) and morality, both at the individual level and in a social context.

Peter Railton (University of Michigan, Philosophy)
Machine Morality: Building or Learning?

The question of how we are to insure that artificially-intelligent systems such as autonomous robots behave in ways consonant with ethical values and norms is often posed, “How to build ethics into robots?” However, impressive progress has been made in artificial intelligence by looking for similarities to how the living brain does things. And the upshot of much of this work is that generic hierarchical *learning* processes are a great deal more powerful than most of us imagined—we need not “build” perceptual features, causal models, or grammars into these artificial systems, since they seem capable of learning most of the content and structure of such organizing schemes. Indeed, they seem of adaptively learning how to learn better. This dovetails with recent work in developmental psychology arguing that human infants likewise acquire causal and linguistic models—even theory of mind—via general-purpose Bayesian learning processes, without needing special-purpose, built-in “modules”. Finally, work in behavioral and affective neuroscience indicates that the mind learns *evaluative* representations that are used prospectively to simulate and assess possible actions and outcomes. A crucial role is played in these processes by *empathy*, whether with one’s future self or with others. We thus are beginning to get a picture of how the pieces might be in place for *moral learning* to take place—the acquisition and use in action-guidance of non-egocentric evaluative models of situations, actions, and outcomes. This picture then suggests that we should be asking, not “How to build ethics into robots?” but rather “How to build robots with a capacity for this sort of deep moral learning?—What ‘priors’ and capacities to respond to experience are needed? What kinds of experience are needed to them to learn to be more reliable at moral evaluation and action-guidance?” We may be a long way from full implementation of moral learning by machines, but with the increasing role of autonomous or semi-autonomous artificial systems in our lives, we are already past the point at which we can ignore the need for partial implementation. We, and the machines, then can learn to do better.

ARTIFICIAL INTELLIGENCE AND HUMAN VALUES

Stuart Russell (UC Berkeley, Computer Science)
Provably Beneficial AI

I will discuss the need for a fundamental reorientation of the field of AI towards provably beneficial systems. This need has been disputed by some, and I will consider their arguments. I will also discuss the technical challenges involved and some promising initial results.

Eliezer Yudkowsky (Machine Intelligence Research Institute)
Difficulties of AGI Alignment

Popular discourse has focused on the question of who will control AGI; a moot point unless the AGI alignment problem is solved first. I discuss the fundamental reasons to suspect that this problem will be hard, even given an intuitively adequate choice of task for the AGI. Our personal and cosmological fates may depend on whether somebody can solve certain technical safety problems, much more than it depends on who is standing nearest the AGI. Strategic implications are, first, that we need to look more in this direction and develop a sustainable ecology of technical proposals and technical critique, and second, that arms races around AGI are extremely bad and should be avoided at almost any cost.

Meia Chita-Tegmark (Future of Life Institute) and Max Tegmark (MIT, Future of Life Institute)
What We Should Want: Physics and Psychology Perspectives

A common question regarding future advanced AI is how to align its values with ours or how to ensure that it does what we want. But what do we want? In so far as we can influence the goals of our descendants, what do we want to want? We provide two contrasting perspectives on this question, from the vantage points of physics and psychology, respectively

Wendell Wallach (Yale, Bioethics)
Moral Machines: From Machine Ethics to Value Alignment

As implementing sensitivity to norms, laws, and human values into computational systems transitions from philosophical reflection to an actual engineering challenge, there appears to be a concerted effort to discard the languages of ethics, in favor of a focus on value alignment. This has parallels to discarding the word "soul" in favor of the word "self" during the enlightenment era. The failures of moral philosophy are many, including the inability to forge a clear action procedure (algorithm) for making moral judgments. To the extent that engineers favor moral analysis they usually turn to utilitarianism. The machines they are building are also natural-born stoics. Nevertheless, there does appear to be some appreciation of moral psychology within the AI community. A few AI theorists have noted that logic without emotions can lead to cognitive biases and an inability to be fully sensitive to human values. But before we throw out the languages of ethics, it would be useful to reflect upon what is lost. Perhaps the languages of ethics serve some more valuable functions than merely being a form of politics by other means.

Steve Petersen (Niagara University, Philosophy)
Superintelligence as Superethical

Nick Bostrom's book *Superintelligence* outlines a frightening but realistic scenario for human extinction: genuine artificial intelligence is likely to bootstrap itself into superintelligence, and thereby become ideally effective at achieving its goals. Human-friendly goals seem too abstract to be pre-programmed with any confidence, and if the superintelligence's goals are not explicitly favorable toward humans, it will extinguish us—not through any malice, but simply because it will want our resources for its own purposes. I try to moderate this worry by showing that by Bostrom's own lights, ethical superintelligence is more probable than he allows. First, Bostrom suggests that because we cannot hardwire a final goal of any real complexity, a superintelligence must *learn* what its own final goals are. This means in effect that such a superintelligence must *reason* about its final goals, and to reason about one's values opens a door for ethics. Bostrom also believes that because software intelligences would be able to self-duplicate, swap memories with others, and switch hardware easily, there will be no sharp lines between one software intelligence and another. Thus it will be clear to a superintelligence (as it reasons out its own goals) that there are no sharp lines between its goals and the goals of others, which will incline the reasoning toward impartiality. But reasoning about final goals while respecting the goals of others plausibly just is ethical reasoning. And since a superintelligence would be especially good at such reasoning, it would be superethical.

MORAL STATUS OF AI SYSTEMS

Eric Schwitzgebel (UC Riverside, Philosophy) and Mara Garza (UC Riverside, Philosophy)
The Rights of Artificial Intelligences

There are possible artificially intelligent beings who do not differ in any morally relevant respect from human beings. Such possible beings would deserve moral consideration similar to that of human beings. Our duties to them would not be appreciably reduced by the fact that they are non-human, nor by the fact that they owe their existence to us. Indeed, if they owe their existence to us, we would likely have additional moral obligations to them that we don't ordinarily owe to human strangers – obligations similar to those of parent to child or god to creature. Given our moral obligations to such AIs, two principles for ethical AI design recommend themselves: (1) design AIs that tend to provoke reactions from users that accurately reflect the AIs' real moral status, and (2) avoid designing AIs whose moral status is unclear. We also discuss the ethics of creating cheerfully suicidal AIs who want nothing more than to die for our benefit.

John Basl (Northeastern, Philosophy) and Ronald Sandler (Northeastern, Philosophy)
The Case for AI Research Subjects Oversight Committees

While it is widely acknowledged that robust AI would be directly morally considerable, how best to protect the interests of such AI raises difficulties. Given the rapid rate of technological advance in the area of AI as well as the plurality of methods that might be used to develop AI, there is significant uncertainty about the form that AI will take, whether we'll be able to tell when robust AI has been achieved, and what sorts of treatment is appropriate for such research subjects. In light of these uncertainties and the potential for mistreatment if robust AI goes unrecognized, it is imperative that we develop ethically sensitive research oversight. We identify what we take to be the primary ethical aim of AI subjects oversight, to ensure that created AIs are treated in a way that is commensurate with their moral status. We then articulate the various obstacles to achieving that aim. Finally, begin the task of specifying how to constitute, task, and empower an oversight committee to best overcome these obstacles and achieve the ethical aim. In other words, which individuals with what expertise should form the committee? What guidance should they be given in evaluating research protocols? What powers should they be given to ensure the protection of research subjects while allowing researchers to advance the field? We begin to answer each of these questions.

S. Matthew Liao (NYU, Bioethics)
Artificial Intelligence and Moral Status

As AIs gain greater capacities, cognitive and otherwise, it seems reasonable to think that they will also acquire greater moral status. To see whether they do in fact acquire such moral status and what kind of moral status they would acquire, we need a theory of moral status. In this talk, I shall lay out a framework for assessing the moral status of various kind of AIs. I am particularly interested in whether and when AIs can have the same kind of moral status as human beings, that is, whether AIs can be rightholders. I shall also discuss whether AIs can have greater moral status than human beings. Lastly, (if time permits,) I shall also explore the most likely paths by which AIs can acquire human-level moral status.