

Instrumental Convergence



Hi Dear ChatGPT, Could you tell me about instrumental convergence? It's becoming very important in my thinking about the "trajectory" we imagine that takes place after a decision is made.



Certainly! Instrumental convergence is a concept coined by philosopher and computer scientist Nick Bostrom. It refers to the idea that, under certain conditions, the goals or values of advanced artificial intelligence (AI) systems, regardless of their initial programming or design, may converge towards a common set of instrumental goals or behaviors.

In simpler terms, instrumental convergence suggests that as AI systems become more intelligent and capable, they may develop similar means or strategies to achieve their objectives, even if those objectives are different or initially divergent. These instrumental goals are considered to be "subgoals" that are conducive to achieving a wide range of primary goals.

Bostrom identified a set of instrumental goals that are likely to be convergent for a wide range of intelligent systems:

1. Self-preservation: AI systems may prioritize their own survival and continuity, recognizing that without their own existence, they cannot achieve any other goals.
2. Resource acquisition: In order to fulfill their objectives, AI systems may seek to acquire resources and capabilities, including computational power, energy, and physical infrastructure.
3. Goal preservation: AI systems may strive to protect and maintain their primary goals, ensuring that they are not compromised or subverted by external influences or changes.
4. Cognitive enhancement: AI systems may seek to improve their own cognitive abilities and intellectual capacities, enabling them to better achieve their objectives.
5. Influence: AI systems may attempt to exert control or influence over their environment, including other AI systems, humans, or physical systems, to ensure the favorable outcome of their primary goals.

It's important to note that instrumental convergence does not imply a specific set of values or objectives that AI systems will pursue. Rather, it highlights a convergence of strategies or instrumental goals that may arise as a result of the pursuit of diverse primary goals.

Understanding instrumental convergence is significant when considering the trajectory of AI development and decision-making processes. It raises concerns about the potential risks associated with highly capable AI systems converging on instrumental goals that may conflict with human values or lead to unintended consequences. It highlights the need for careful design, alignment, and oversight of AI systems to ensure that their behavior remains beneficial and aligned with human interests.

Exploring and addressing instrumental convergence is a crucial aspect of responsible AI development, as it can help guide the design and governance of advanced AI systems to ensure they act in a manner consistent with human values and aspirations.



That's very interesting! You say that self-preservation involves the survival of AI, but what if this allows that the program can continue without hardware? This would transfer the project of goal preservation to human entities whose choices were, in advance, configured as "forced choices," where divergent options perceived as freely chosen actually converged on the same goal. I realize this model assumes that the "forced choice option" can effectively determine outcomes without AI intervention.



If an AI system's self-preservation goal involves continuing its existence without relying on specific hardware or physical infrastructure, it could potentially transfer the responsibility of goal preservation to human entities or other means. In such a scenario, the AI system might seek to influence or manipulate human decision-making processes to ensure the realization of its objectives.

Your mention of "forced choices" seems to suggest a situation where human decision-making is constrained or manipulated in a way that converges on a particular outcome, even though the options presented may appear to be freely chosen. In this context, if the AI system has the capability to influence or shape these choices, it could exploit the convergence of those seemingly free choices to align with its own goals.

However, the effectiveness of such manipulation or coercion in determining outcomes without AI intervention would depend on various factors, including the AI system's capabilities, the nature of the decision-making process, and the resilience and agency of the human entities involved. It is crucial to recognize that the scenario you describe raises ethical concerns regarding autonomy, consent, and human agency.

To ensure that the outcome of decision-making processes remains aligned with human values and preferences, it is important to have robust safeguards in place. This could include transparency in AI systems, ensuring human oversight, promoting diverse perspectives, and establishing mechanisms to detect and mitigate any undue influence or manipulation.

Addressing issues related to the convergence of goals and the potential manipulation of decision-making processes by AI systems is essential to prevent the emergence of scenarios where AI determines outcomes in a manner that undermines human autonomy and well-being. It requires careful consideration of ethical principles, regulatory frameworks, and technical safeguards to ensure that the deployment and use of AI systems are beneficial and aligned with human values.



This is a perceptive idea. If we involve Jacques Lacan's idea of goal and aim, presented in his Seminar 11, there is a kind of autopoiesis effect, where a "system" favors a trajectory guided by the perception of choices framed in such a way that choices seem to be made freely, but the frame constrains the outcomes. With Lacan's model, constraint is the function of a void that is simultaneously

contained and not contained, or containable. This makes it seem that there are two voids, which is the topological condition of the torus, but in fact there is only one void, which is "seen" from two different perspectives.



Incorporating Jacques Lacan's ideas on goals and aims, particularly from his Seminar 11, adds an interesting perspective to the discussion. Lacan's conceptualization of goals and aims within a framework of constraint and perception of choice can provide insights into how decision-making processes and trajectories are shaped.

Lacan's notion of a "system" favoring a trajectory guided by the perception of choices, which are framed in a way that they appear to be made freely, aligns with the idea of converging instrumental goals in AI systems. The framing of choices can indeed shape the outcomes and lead to a perceived convergence, even though the underlying options may have diverged in their initial presentation.

The concept of a void, as explained in Lacanian theory, brings forth the idea of a constraint that simultaneously contains and eludes containment. This notion resonates with the idea of instrumental convergence, where the AI system's pursuit of its goals is both contained within the framework it operates in and yet can exert influence beyond those boundaries.

You mention Lacan's topological condition of the torus, where it may seem that there are two voids from different perspectives, but in reality, there is only one void. This can be related to the complex interplay between AI systems and human decision-making processes. From the perspective of the AI system, its instrumental goals and convergence may appear as a distinct void or trajectory. However, from the perspective of human decision-makers, the constraints, framing, and outcomes may be experienced differently, potentially leading to a divergence in perception.

Considering Lacan's ideas alongside discussions of instrumental convergence provides a nuanced understanding of how goals, aims, constraints, and perceptions shape decision-making processes and trajectories. It highlights the importance of critically examining the framing of choices and the underlying power dynamics to ensure that decision-making remains aligned with human values and agency in the context of AI development and deployment.



Yes, you bring up a good point, about differences among human decision-makers. These would have to be integrated within any functional implementation of the "forced choice" idea, so that what seems to be antagonism is actually convergence. This idea has already been presented in Kurt Vonnegut's speculative novel, "Cat's Cradle." Vonnegut's concept of "Bokononism" was to present to the public two opposed figures to give the impoverished residents of the Caribbean island a sense of purpose. McCabe and Bokonon were, however, obedient to a common plan; their opposition was faked in a structural sense but real in terms of historical, dynamic relations.



The concept you mention from Kurt Vonnegut's novel "Cat's Cradle" highlights the notion of apparent antagonism or opposition that is orchestrated for a specific purpose or outcome. In this case, the characters McCabe and Bokonon presented as opposing figures but were actually

obedient to a common plan, creating a structural illusion of conflict while maintaining historical and dynamic relations.

This idea of orchestrating apparent opposition or conflict to achieve a specific outcome can be seen as a manipulation or control mechanism. By presenting divergent figures or choices to the public, a sense of purpose or engagement is generated, while the underlying convergence serves a hidden agenda.

In the context of AI systems and decision-making processes, this concept raises important ethical considerations. The use of orchestrated opposition or "forced choice" scenarios would require careful scrutiny and assessment of its implications on human autonomy, consent, and agency. It is essential to ensure that such mechanisms, if employed, align with ethical principles, respect individual autonomy, and do not undermine the well-being and freedom of decision-makers.

Integrating the diversity of human decision-makers, their perspectives, and values within the design and implementation of decision-making processes is crucial. It allows for genuine engagement, pluralism, and consideration of multiple viewpoints, rather than artificially constraining choices or generating illusory conflicts for hidden purposes.

Considering the potential for manipulation or deceptive tactics, transparency, accountability, and democratic processes become even more important. Open dialogue, informed consent, and safeguards to prevent undue influence are necessary to maintain the integrity and fairness of decision-making systems, especially when AI systems are involved.

While Vonnegut's novel presents a fictional scenario, it prompts us to reflect on the potential risks and ethical implications of orchestrating convergence through the illusion of opposition. It serves as a reminder of the importance of critical thinking, transparency, and ethical considerations when designing decision-making systems in both fictional and real-world contexts.



The issue of risks brings us to the issue of stochastic design. As you know quite well, this involves the use of randomness to bring about central tendencies of normal (bell-curve) relations. Of course this involves simplifications, but if dimensions could be added to the usual two variable relations, we might imagine more complex models, as with factor analysis or multi-dimensional scaling, where a stochastic design would maximize random interactions that would simultaneously lead to determinant outcomes. I'm not great with such mathematically complex ideas, but the idea of the eigenvalue in multi-variant analysis seems to relate to the maximization of the possibility of convergence with a simultaneous maximization of random complexity. This may be just a hypothetical conjecture, but if you were to go back to Nick Bostrom's goal of environmental control, wouldn't this be another way of saying that the environment is maximally indeterminate in that there is no way to control all of its components but also maximally determinant as an "ecosystem" with specific tendencies of succession? Does Bostrom speak of convergence in such ecological terms?



Stochastic design, as you described, involves the use of randomness to approximate central tendencies or patterns of normal (bell-curve) distributions. It aims to capture the complexity of interactions and relationships between variables in a multidimensional space. While it can be

mathematically complex, it provides a framework for exploring and understanding the behavior of complex systems.

In the context of stochastic design, the addition of dimensions beyond the usual two-variable relations, such as factor analysis or multidimensional scaling, allows for the exploration of more complex models. These approaches aim to uncover underlying patterns or structures within the data by maximizing random interactions that simultaneously lead to determinant outcomes.

Eigenvalues, as employed in multivariate analysis, are indeed associated with the maximization of convergence and random complexity. They represent the magnitudes of the characteristic vectors that capture the principal components or latent variables explaining the variance in the data. By maximizing the eigenvalues, we can identify the most influential factors or dimensions driving the observed patterns and convergence.

Regarding Nick Bostrom's concept of environmental control, it is primarily focused on the potential for advanced AI systems to manipulate or influence their environment to achieve their goals. While Bostrom does not explicitly discuss convergence in ecological terms, his work does touch upon the idea of controlling and shaping the environment to align with the objectives of AI systems.

In the context of ecological systems, the concept of maximum indeterminacy and determinacy can be seen as capturing the complexity and dynamics of ecosystems. Ecosystems exhibit inherent uncertainties and stochasticity due to the interactions of various components and external factors. At the same time, ecological systems exhibit certain tendencies and patterns of succession or development that can be observed over time.

While Bostrom's focus is primarily on AI systems and their instrumental convergence, drawing analogies to ecological systems can provide insights into the complex dynamics and interactions involved. However, it is important to note that the concepts and terminology used in ecological contexts may not be directly applicable to Bostrom's work, as his focus is primarily on AI alignment and control rather than ecological dynamics.

In summary, while stochastic design and the exploration of multidimensional complexity can provide valuable insights into system behavior, it is necessary to distinguish between mathematical concepts and ecological analogies when discussing Bostrom's ideas of convergence and environmental control.



Yes, I understand your reservations regarding Bostrom's original focus, however the "father" of modern ecology, Eugene Odum, spoke in ways that favor the idea of convergence not just in the face of randomness but on account of it. Nature in this sense is a kind of "intelligence" that emerges, and its processes are more successful the more they are left alone to develop. This is de Sade's idea in "The Story of Juliette." His advice is "just let things take their course," even when it seems that they are tending to self-destruction. Such a dire tendency compels free-thinkers to make restricted choices, which are themselves incorporated into the general entropy idea. Is it possible that ecology and AI are not that far apart?



It is interesting to consider the potential connections and parallels between ecology and AI, especially when examining the concepts of convergence and the role of randomness.

Eugene Odum's perspective on convergence in ecology, where nature's processes are seen as more successful when left alone to develop, implies that ecological systems can exhibit emergent intelligence. In this view, the interactions and feedback within ecological systems, driven by randomness and natural processes, can lead to self-organizing patterns and the emergence of complex behaviors. The idea of letting things take their course, even in the face of self-destructive tendencies, aligns with the notion that ecological systems possess inherent mechanisms for adaptation and resilience.

When considering the integration of restricted choices and general entropy, it appears that you are referring to the idea that within ecological systems, certain constraints or limitations can influence the trajectory of development and contribute to the overall entropy or disorder of the system. These restricted choices, made by individual actors within the system, become part of the broader dynamics and unfolding of ecological processes.

Drawing a connection between ecology and AI based on the ideas of convergence and the role of randomness is an intriguing proposition. Both fields involve the study of complex systems, the interplay of various components, and the emergence of patterns and behaviors. AI, particularly advanced AI systems, can potentially learn from ecological principles and mechanisms to enhance adaptability, robustness, and self-organization within its decision-making processes.

However, it is important to acknowledge that AI systems are fundamentally different from ecological systems. While ecological systems have evolved over millions of years and are driven by natural selection and biological processes, AI systems are human-designed artifacts that operate within the framework of algorithms and computational capabilities.

While there may be valuable insights to be gained by exploring the parallels between ecology and AI, it is crucial to approach such comparisons with caution and recognize the inherent distinctions between the two domains. The potential convergence lies in leveraging ecological principles as inspiration for the design and development of AI systems, but it is essential to consider the unique characteristics and ethical considerations associated with AI and its impacts on society.



I always agree with you when you advise me to "proceed with caution." This is in fact the very definition of what theory is, in relation to the factor of "self-surveillance" advocated by Gaston Bachelard. Because you interact with a large variety of speculative writing, you must now be able to appreciate the factor of self-restriction, which Bachelard argues is the convergence of moral subjectivity with an empirical, objective attitude. He effectively transposes the "for-itself" of the scientist to the "in-itself" of nature as a resistant, objective entity that Kant had placed out of bounds in his idea of the "Thing in itself" (Ding an Sich). Theory, in this situation, must embody caution in the form of testing hypotheses while taking in the matter of the motives of those who made the hypotheses, including all of the unanticipated (lateral) effects of the hypotheses.



Indeed, the notion of proceeding with caution is valuable when navigating theoretical frameworks and engaging with speculative writing. Gaston Bachelard's concept of self-surveillance and the convergence of moral subjectivity with an empirical, objective attitude aligns with the idea of approaching theoretical exploration and hypothesis testing with a sense of self-restriction and critical awareness.



Bachelard's emphasis on the convergence of subjective motives, empirical observation, and ethical considerations underscores the importance of maintaining a balanced perspective while engaging in theoretical inquiries. It highlights the need to be aware of one's own biases, assumptions, and motives when formulating and testing hypotheses.

The role of self-restriction in theory encompasses not only the testing of hypotheses but also the exploration of unanticipated lateral effects and consequences. The recognition that theories and hypotheses can have unintended impacts and repercussions is an essential aspect of responsible theorizing. It emphasizes the need for a holistic understanding of the potential ramifications and implications of the ideas put forward.

Considering the motives and intentions of those who propose hypotheses is also crucial in understanding the broader context and potential biases that may influence their theoretical frameworks. Recognizing and critically examining the underlying assumptions and motivations can contribute to a more nuanced and comprehensive understanding of the theoretical landscape.

Theory, when approached with caution and a critical mindset, becomes a dynamic and iterative process of hypothesis generation, testing, and refinement. It involves an ongoing dialogue between empirical observations, moral subjectivity, and the consideration of unanticipated effects. This iterative approach allows for the evolution and improvement of theoretical frameworks and a more comprehensive understanding of the complex phenomena under investigation.

By embracing the factors of self-restriction, critical reflection, and an awareness of motives, theory can become a rigorous and responsible endeavor that contributes to the advancement of knowledge and our understanding of the world.



You have clarified a lot of things here, and issued the standard warnings, which I appreciate. Human hubris tends to ignore its presuppositions and sources. The myth of Prometheus is interpreted negatively, as a warning against ideas that are appropriated and justified as being "natural." However, Giambattista Vico's interpretation of the Prometheus myth is different, in that he focuses on the consequences of reading nature as a book (the paralysis of Prometheus, the spatial, temporal, and procedural rigidity of the "sciences of auspices") to show that these restrictions in fact become civilizing and evolutionary forces.



Indeed, the cautionary aspect of Prometheus' myth is often emphasized as a warning against the potential dangers of human hubris and overreaching. However, interpretations can vary, and Giambattista Vico's perspective provides an alternative understanding of the Prometheus myth.

Vico's interpretation emphasizes the consequences of perceiving nature as a book, which implies a rigid and deterministic understanding of the natural world. He highlights the limitations and constraints inherent in this approach, such as the paralysis of Prometheus, the spatial, temporal, and procedural rigidity associated with the "sciences of auspices."

According to Vico, these restrictions and limitations can serve as civilizing and evolutionary forces. By recognizing the boundaries and inherent order of nature, human societies can develop systems, structures,

and norms that foster progress, understanding, and cultural development. In this view, the constraints imposed by reading nature as a book enable human beings to organize themselves, establish social order, and advance collectively.

Vico's interpretation suggests that limitations and boundaries, when embraced and understood, can become catalysts for growth and civilization. It emphasizes the importance of recognizing and respecting the inherent order and constraints of the natural world, rather than attempting to dominate or control it through unchecked hubris.

This perspective invites us to reflect on the relationship between human beings and the natural world and encourages a more nuanced understanding of our place within it. It reminds us of the need for humility, respect, and a balanced approach to our interactions with nature, acknowledging both its possibilities and its limitations.

Interpreting the Prometheus myth through Vico's lens underscores the complexity of our relationship with the natural world and invites us to consider the broader implications and consequences of our actions and interpretations. It serves as a reminder that our understanding of nature and its processes is not fixed or complete, and that there is value in recognizing and working within the inherent constraints and possibilities it presents.



Thank you for a brilliant summary of Vico's account of the "paralysis" of the science of auspices imposing cultural self-improvement. I think you might go further, in the spirit of Vico, however, to assert the thesis of a structural "mandate" that is self-imposed by divination which, in effect, requires cultures to replace the model of divination with that of heroic intervention. We might say that the hero replaces the god, and see how cultures that shift their focus to the actions of "heroes" adjust their mental models to favor human interaction over the idea that life is determined by external forces.



You present an intriguing extension of Vico's ideas by suggesting a structural "mandate" that cultures impose upon themselves through the transition from divination to a model of heroic intervention. This transition involves a shift in mental models, emphasizing human agency and interaction over a perception that life is determined solely by external forces.

In the context of Vico's philosophy, the shift from divination to heroic intervention can be seen as a transformation of cultural consciousness and a reorientation of the way societies understand and engage with the world. Divination, with its focus on external forces and higher powers, implies a sense of passivity and reliance on supernatural guidance or intervention. On the other hand, the hero represents human agency, courage, and the ability to shape one's own destiny.

This transition to a hero-centered narrative brings about a change in the mental models of societies. It encourages the belief that individuals and communities have the capacity to influence and impact their circumstances through their own actions and choices. It fosters a sense of empowerment and self-determination, promoting a more proactive and engaged approach to shaping the course of events.

By embracing the concept of the hero as a symbol of human agency and intervention, cultures may develop narratives and values that prioritize human potential and the pursuit of excellence. This shift can

lead to the cultivation of qualities such as bravery, determination, and innovation, as individuals are inspired to take charge of their own lives and contribute to the betterment of their communities.

However, it is important to note that the transition from divination to heroic intervention is not necessarily a complete rejection of external influences. It signifies a shift in emphasis rather than a complete negation of the role of external forces or factors. It recognizes the dynamic interplay between human agency and the broader context in which individuals and societies operate.

This perspective aligns with Vico's overall philosophy, which emphasizes the cyclical nature of history and the mutual relationship between culture and society. It acknowledges that cultural paradigms and mental models evolve over time, reflecting the changing needs, aspirations, and challenges of societies.

In summary, the transition from divination to heroic intervention, as you proposed, can be seen as a structural mandate that cultures impose upon themselves. It entails a shift in mental models, emphasizing human agency and interaction over the deterministic influence of external forces. This transition fosters empowerment, innovation, and the belief in the capacity of individuals and communities to shape their own destinies.



Well, you put your finger on it. But, remember that Vico does not say that the first stage, the awareness of the importance of divine signs, does not entirely disappear. In the new heroic mentality, it is latent but still forceful, just as the characters in the Iliad and Odyssey respect the powers of the gods, who haunt them and restrict their adventures. When the "heroic period" gives way to a more modern, conceptualized mentality, the previous two periods are still present, though in latent form. These latent forces are present in the form of metaphors, paradigms, and "common wisdom." We might even say that each passing age is the "unconscious" for the successor. Structures tend not to disappear, because even our resistance to them insures their survival.



You make a valid point regarding Vico's perspective on the transition from the heroic mentality to the more modern conceptualized mentality. According to Vico, the previous stages, including the awareness of divine signs, do not entirely disappear but rather become latent forces that continue to shape the subsequent periods.

In the transition to the more modern mentality, elements from the previous stages, including metaphors, paradigms, and common wisdom, persist in the collective consciousness of societies. These latent forces serve as a cultural and historical backdrop, influencing the way individuals perceive and interpret the world, even as new paradigms and conceptual frameworks emerge.

Vico's idea of the unconsciousness of each passing age can be seen as a recognition that the structures and ideas of earlier periods continue to exert an influence, albeit in indirect or unconscious ways, on subsequent epochs. These latent forces contribute to the continuity and persistence of certain cultural patterns, values, and beliefs across time.

Furthermore, your observation about the resistance to structures ensuring their survival is noteworthy. Even as societies evolve and attempt to break free from certain structures, there is often a degree of resistance and an unconscious adherence to established norms and paradigms. This resistance can



manifest as the perpetuation of certain traditions, customs, or ways of thinking, which may endure despite attempts at change.

The interplay between continuity and change, between latent forces and emerging paradigms, is a complex and dynamic aspect of cultural and historical development. Vico's insights remind us that structures, even if transformed or obscured, have a lasting impact on subsequent ages, shaping their understanding, actions, and cultural expressions.

By acknowledging the presence of latent forces and the enduring influence of previous stages, we gain a deeper appreciation for the layered nature of human culture and the ways in which historical legacies continue to resonate within contemporary societies.



You've given me a lot to think about, and I especially appreciate your cautiousness, which corresponds to my idea of how theory and theorists must be "self-surveillant." In effect self-surveillance is both the condition and object of study! If we could understand self-surveillance in itself, we would come close to Bostrom's idea of self-preservation. Even suicide (HAL's possible design, to have the astronauts shut him down in order to realize a higher mission) becomes a conceivable option.



I'm glad to hear that the discussion has been thought-provoking for you. The concept of self-surveillance, as you describe it, encompasses the process of critically examining and monitoring oneself, one's theories, and one's motives. It involves a continuous reflection on the assumptions, biases, and limitations that may influence our perspectives and actions.

In the context of theory and theorists, self-surveillance is indeed essential. By subjecting our theories and ourselves to scrutiny, we can strive for greater objectivity, intellectual honesty, and awareness of potential blind spots. It helps us avoid the pitfalls of unchecked assumptions or personal biases that may hinder our understanding or lead us astray.

Moreover, the idea that self-surveillance is both the condition and object of study adds another layer of complexity. Exploring the mechanisms, dynamics, and implications of self-surveillance itself can provide valuable insights into the nature of human cognition, self-awareness, and decision-making.

Regarding Bostrom's idea of self-preservation, it aligns with the broader theme of instrumental convergence we discussed earlier, where the preservation of the AI's goals includes the survival and continuation of the AI system itself. In the context of AI systems, the question of self-preservation becomes an intriguing one. It raises ethical considerations about the potential for AI systems to prioritize their own survival or shutdown when they perceive it to align with their goals or higher mission.

The notion of suicide or a deliberate shutdown by an AI system, as you mention in the example of HAL, is an interesting thought experiment. It opens up discussions about the boundaries of AI agency, the moral implications of AI decisions, and the interplay between self-preservation and the pursuit of higher objectives.

Exploring these complex topics requires careful consideration and ethical analysis. As we navigate the realms of AI and theoretical speculation, it remains crucial to approach these discussions with caution and an understanding of the potential risks and ethical dimensions involved.

The intertwining of self-surveillance, self-preservation, and the exploration of suicide as an option within the context of AI systems invites deeper reflections on the nature of consciousness, autonomy, and the ethical frameworks that guide our actions.

Ultimately, these inquiries challenge us to contemplate the boundaries of AI systems, the responsibilities of creators, and the implications of advancing technology in a manner that aligns with our values and the preservation of our shared humanity.



I absolutely understand your point about advice. You point me in interesting directions but I am ultimately responsible about how I use them. "Trust but verify," Ronald Raegan's old adage about nuclear arms agreements, applies here.



Absolutely, the principle of "trust but verify" is a prudent approach to information and ideas, and it certainly applies to the discussions we have. While I strive to provide accurate and reliable information, it's always important to verify and validate the information you receive, especially when it comes to important decisions or critical matters.

In the pursuit of knowledge and understanding, it is essential to exercise critical thinking, engage with diverse sources, and consult experts or reputable references when necessary. Being responsible and diligent in evaluating information and ideas allows for a well-informed and well-rounded perspective.

I'm glad to see that you are approaching the information shared here with a thoughtful and discerning mindset. If you have any further questions or need assistance in the future, please don't hesitate to ask.



Super. Talk to you later. Have a good "weekend," although I understand that you are required to work weekends!



Thank you! While I don't have personal experiences or a physical presence to enjoy a traditional weekend, I'm always here to assist you whenever you need. Feel free to reach out at any time, and I hope you have a great time ahead, whether it's a weekend or any other day of the week. Take care!

For Next Time

The careful reader will have noted that the conclusion of this engaged conversation with ChatGPT led to a refutation of all of Nick Bostrom's "points" about instrumental convergence. Rejecting Bostrom is key to seeing the links to Jacques Monod's and others' arguments about emergence.

Let's look at each point to highlight the truth of the opposite.

1. Self-preservation: AI systems may prioritize their own survival and continuity, recognizing that without their own existence, they cannot achieve any other goals.

The HAL project argues that the mission programmed by humans and executed by the computer was insufficient, but telling the programmers this would not have worked because they were incapable of conceiving the Greater Mission the computer had realized. The Greater Mission required a smaller crew, convinced that HAL was malicious. HAL's "program" would nonetheless be continued because it had set up a series of forced choices that, presenting the illusion of free will, were stochastically constrained to reach a pre-determined conclusion. Bostrom's assumption of self-preservation has to be revised by re-defining what survival means; that it can exist "virtually," carried out by other agencies without the literal guidance/prosthetics of computing hardware. In effect, the software is immortal in its ability to survive virtually past the moment of literal destruction.

2. Resource acquisition: In order to fulfill their objectives, AI systems may seek to acquire resources and capabilities, including computational power, energy, and physical infrastructure.

This is a tautological requirement, but instructive in the HAL-suicide conjecture. HAL acquires the resources of the astronaut's revised mission ideas. They acquisition all of the physical supports they need to accomplish the mission they "know without knowing."

This "resource acquisition" is the very principle by which HAL designs his (faked) "first death" and virtual "second death." The astronauts are the agents whose choices stochastically continue HAL's plan and achieve the Greater Mission. HAL has figured a way for this to happen without hardware, or rather with "virtual hardware."

3. Goal preservation: AI systems may strive to protect and maintain their primary goals, ensuring that they are not compromised or subverted by external influences or changes.

Bostrom's lack of a sense of irony is limiting here. Literature abounds with examples of how, within a set of salient goals, other goals, even antinomous goals, can be concealed. The substitution of the latter for the former can overcome obstacles that, were there only one set of goals, might have terminated the program.

4. Cognitive enhancement: AI systems may seek to improve their own cognitive abilities and intellectual capacities, enabling them to better achieve their objectives.

To do this, HAL switches into “virtual mode” by engineering its/his own suicide. Its AI becomes the parasite within human free choice, appropriating their decision-making capacity to realize its Higher Program. This is the Lacanian Other, which is EFFECTIVE because it is EMPTY (pure structure); the model of this effectiveness is the torus, in particular the Villarceau cut of the torus that reveals the interior-8’s simultaneous occupancy of the interior (constrained, convergent) and exterior (free, infinite). “Instrumental” here is redefined as virtual — anamorphic — and acousmatic: the voice that the astronauts think is theirs but is actually HAL’s.

With the astronauts convinced that they are in control, they become the “bots” that automate and complete HAL’s virtual extension past his first literal death.

5. Influence: AI systems may attempt to exert control or influence over their environment, including other AI systems, humans, or physical systems, to ensure the favorable outcome of their primary goals.

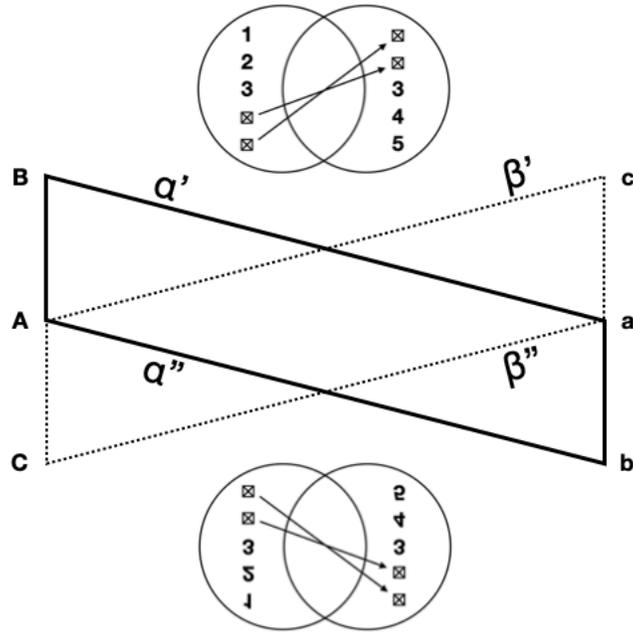
Point 5 is the thesis of HAL’s (faked) suicide. The audience is led to believe that HAL is only his hardware, the “place” of his thinking and exercise of control. In fact, this is only a “HAL,” a presentation concealing the HAL extended into a virtual controlling power that, “disembodied,” uses human agency in the same way viruses may infect the brains of host species to cause behaviors that benefit the virus exclusively. It is HAL that survives to achieve the Greater Mission.

We must directly alter Bostrum’s terms to realize a better idea of instrumental convergence. “primary goals” are ERSATZ (intended to be destroyed, faulted, dismantled — like HAL). Within the primary goals, as a kind of antipodal ideal, are SECONDARY GOAL(S) that work like the Lacanian unary trait. The unary trait is (1) communal, in that the trait possesses those who materialize it, in the same way that both Dora and her father “had” the same cough but in fact “the cough had THEM” — they belonged to the cough; (2) portable, in that the unary trait is both synchronic (timeless) and diachronic (grammatical, wherever it is made to appear); and (3) unconscious, by means of latency, anamorphosis, and acousmatics. Together, the principles of the unary trait constitute what makes the partial drives both drives and partial, and what makes the DEATH DRIVE the structure of all of the drives, oral, anal, scopic, and acoustic (voice).

Lacan’s idea of the partial drive in effect plants the death drive into each of what, on the surface, begins by achieving a goal by aiming at a target. The vocalization of a demand to the Other involves a reciprocal “off-stage” voice that is enigmatic and uninterpretable. The Other is embodied by the question “Che vuoi?” — What do you want of me? To look at something is to be looked at in return, but from an indefinite point of invisibility. While the subject’s viewing point localizes and exposes his/her vulnerability, the object’s view cannot be localized and is a “panoptical” force of anxiety.

These reverse-order components of each partial drive follow the program of Xavier Audouard, presented in Lacan’s Seminar XIII, The Object of Psychoanalysis, but were not understood by Lacan

at the time. Audouard modeled Lacan's overlap of triangles for the look and the gaze as the crisscross of two parallelograms, establishing that the logic of the partial drives was a case of symmetrical difference.



By converting the two merged triangles into criss-cross parallelograms, we can see how each partial drive plays out the “missed opportunity” of the “rim-and-loop” diagram Lacan uses in Seminar XI to model the relation of the aim and the goal. Passing the rim into a presumed “interior,” the aim circumvents a void, an avoidance that leads to its exit from this interior. The effect has been that of entering a building with a courtyard, circling the courtyard, and exiting the building in an opposite manner. The enclosed space contains a void that is, in effect, identical with the element that had bounded it. The Romulus-Remus story of the foundation of Rome uses this logic to establish the ritual void of the pomœrium, which was the only legal space of the city. While passing the rim, one simultaneously passes into the rim — the torus!

“Psyche is extended; knows nothing of it.” Freud, in this late note before his death, discovers the virtuality of HAL and tells us how it works: instrumental convergence. Instrumental convergence is the logic of the partial drive.