Humanity and Heuristics: How Ethical Programming Can Ensure the Future Symbiosis of Man and Machine

160026981
November 13th, 2020
Word Count: 3,431

> *"Everything we love about civilization is a product of intelligence, so amplifying our human intelligence with artificial intelligence has the potential of helping civilization flourish like never before – as long as we manage to keep the technology beneficial."*
>
> - *Max Tegmark, President of the Future of Life Institute*

Introduction

"I'm sorry Dave, I'm afraid I can't do that." – HAL 9000 (2001: A Space Odyssey)

Artificial Intelligence has been portrayed in science fiction pop culture dating back to 1927.[1] Since then, AI has been presented on the big screen in a variety of ways – from the symbiotic companion of human heroes, to apocalypse inducing races of robots bent on human destruction. While *Terminator* style killer robots paint a popular dystopic future for AI and Humanity, perhaps the most resonant representation of human-AI conflict comes from the 1968 film *2001: A Space Odyssey* by Stanley Kubrick.

The film's artificial antagonist HAL 9000 (Heuristically Programmed Algorithmic Computer) is an AI agent that manages all of the systems on a spaceship while interacting with the humans in the ship's crew. As the film progresses, HAL begins to malfunction and acts in ways that result in human casualties. The agent's goal driven behavior and corrupted objectives ultimately drive it to make decisions which endanger the humans onboard. When ordered by a human to perform a task which could potentially rescue a crewmate from HAL's fatal actions, HAL famously responds "I'm sorry Dave, I'm afraid I can't do that."[2]

While Kubrick's film is over 50 years old, the concerns over ethical artificial intelligence raised by HAL are in line with scholars today. AI will continue to improve and will become extremely good at achieving its desired goals. The Future of Life institute outlines two potential ways that AI can become dangerous.[3] The first is an intentional programming of weaponized AI. The second is that AI is intended to be beneficial, but inadvertently becomes destructive in its fulfillment of its purpose. This paper will focus on the second threat of AI – unintended consequences of developing super-intelligent systems.

The danger to humanity arises when an agent makes decisions to achieve objectives that undermine human safety and morality. As artificial intelligence technology improves, the ability to create "good" or "evil" AI can be controlled partially by understanding how to align AI goals with humanitarian ethics. By ensuring that AI does not prioritize efficiency over ethics, humans can limit potentially dangerous consequences, and leverage AI for a symbiotic relationship.

---

[1] (Tomlinson, 2018)
[2] (Kubrick, 1968)
[3] (BENEFITS & RISKS OF ARTIFICIAL INTELLIGENCE, n.d.)

Singularity

> *"Success in creating effective AI, could be the biggest event in the history of our civilization. Or the worst. We just don't know. So, we cannot know if we will be infinitely helped by AI, or ignored by it and side-lined, or conceivably destroyed by it,"*
>
> *– Stephen Hawking* [4]

The concept of a technological singularity has been around since the 1960's – first introduced by I.J. Good who coined the idea of an "intelligence explosion" brought forward by an "ultra intelligent machine" that would occur in the twentieth century and ultimately create an ever-growing gap between human and machine intelligence.[5] From Alan Turing to Elon Musk, the concept of singularity has been recognized as the major transition point in Human-AI interaction.[6] The singularity refers to the point in time when artificial intelligence surpasses human intelligence, allowing AI to improve autonomously in an explosion of technological growth and creating a great distance in intelligence between humans and AI in a very short amount of time. [7] The AI singularity, described famously by Vinge in 1993, has the potential to reshape the rules of progress itself. While biological species rely on natural selection for adaptation and innovation, a superintelligence would be able to execute simulations to find solutions and evolutions thousands of times faster than even the highest human intelligence.[8] This theory of recursive improvement by machines could create an environment where a superintelligence emerges superior to the human brain in both computation and creativity.[9]

There are conflicting opinions over the plausibility of a true singularity event in AI. Kurzweil (2005) hypothesized that the exponential growth of computing power and software productivity could result in a singularity event by the middle of the 21st Century. [10] However, while hardware and software improvements are relevant, AI technology has not yet been optimized by humans. Korb and Nicholson (2014) point out that researchers have yet to agree on an optimized AI methodology, or produce a machine capable of passing the Turing Test in 60 years of the field. [11]

In order for a machine to be super intelligent, it must first achieve human-level intelligence. However, Jackson Jr. (2018) theorizes that human-level AI can be achieved without

---

[4] (Kharpal, 2017)
[5] (Good, 1966)
[6] (Walsh, 2017)
[7] (Schrader & Ghosh, 2018)
[8] (Vinge, 1993)
[9] (Burkhardt, 2011)
[10] (Kurzweil, 2005)
[11] (Korb & Nicholson, 2014)

the need to replicate "human-like" intelligence and consciousness, arguing that the Turing Test is not necessarily the correct measurement of human intelligence. Jackson's TalaMind thesis conjectures that if human-level AI is achievable, then weak superintelligence, or improved human intelligence is achievable. [12] However, the path from weak superintelligence to "strong superintelligence" is less clear – relying on presently unknown characteristics of the computational principles of the human brain.[12]

Walsh (2017) outlines some fundamental arguments against the occurrence of a singularity. He notes that at a basic level, there is no evidence that humans have the intelligence to design an artificial intelligence capable of causing a singularity event. Likewise, he introduces the argument that singularity theory confuses the intelligence to perform a task, with the capability to improve one's intelligence to perform the task. [13] In other words, even the most advanced AI algorithms are only improving via the work of humans to improve the algorithm's specific task performance. So, if superintelligence (AI+) is achieved, how can it be guaranteed that the recursive improvements will focus on the agent not only learning to improving task performance but learning to learn better than humans in general.[13] Despite his arguments against a singularity occurrence, Walsh still concludes that super intelligent machines are possible, and regardless the future impact of AI on society should start being understood. [13]

While there are varying perspectives about the true imminence[11] and definition[14] of a singularity, there is equal polarity amongst experts regarding the outcome of a singularity on the human race. As Walsh (2017) notes, there are AI optimists and pessimist when it comes to the singularity – and the views of both make it important to address. [13] In the worst case, super intelligent agents would have different concepts of self-preservation and optimization than humans, resulting in the extinction of the human race.[15] Vinge offers that a scarier alternative than extinction is living as an inferior race to superintelligence, serving as mere signal processors or fuel to the new dominant intelligence. [15] Concern arises over the perspective and motivation of superintelligence – if AI is indifferent or threatened by human civilization it may easily reconfigure the environment, resources and human existence. [16]

In the best case, superintelligence can help humans advance science, medicine, space travel and more to the point where it becomes a critical part of survival. [12] Rather than a true divergence between man and machine, superintelligence combined with human intelligence could form a greater collective intelligence that capitalizes on technological and social advancements to further continual growth for both AI and humanity. [17] Some theorists predict a

[12] (Jackson, Jr., 2018)
[13] (Walsh, 2017)
[14] (Sandberg, 2013)
[15] (Vinge, TECHNOLOGICAL SINGULARITY, 1993)
[16] (Mulgan, 2016)
[17] (Prescott, 2013)

fluidity between man and machine in the future, where bioengineering allows physical integration of AI and the human brain. [18, 19]

The difference between the best-case and worst-case scenario lies in humanity's ability to encode ethical goals within AI. Ensuring that an AI will be able to reason about goals, and the implications of the possible means to achieve them, can help prevent the scenario where an optimization-based machine eliminates human society.[20] While the prospect of a singularity seems like the plot of a science fiction movie, work has already begun on developing ethical, or even super-moral AI in an effort to ensure a beneficial relationship between humans and the intelligence of the future.

Ethical Programming

> *"AI doesn't have to be evil to destroy humanity - if AI has a goal and humanity just happens in the way, it will destroy humanity as a matter of course without even thinking about it, no hard feelings…"*
>
> *-Elon Musk [21]*

In the case of a singularity event, the prominent threat is that a superintelligence will prioritize logic, optimization and utility while ignoring implications to humanity and being void of empathy or intuition.[22] Jackson (2018) proposes that one way to avoid such a scenario is to ensure that any superintelligence can understand the relationship between goals and implications of achieving the goal based on an "ethical meta-goal" – even allowing the agent to abandon or change its goal in the face of harmful consequences to humanity.[20] Building ethical AI is a present industry problem as well. Autonomous vehicles and AI in the medical space will be faced with problems that have ethical decisions to be made.[23] Since 2016 there has been increased interest in AI and robot ethics, seeing participation from leading technology companies and even national governments.[23]

The fear of AI controlling human life may seems distant but is already a growing concern today. Weak AI is already playing a role in the news we see, the products we buy and even the

---

[18] (Goertzel, 2007)
[19] (Nicolescu, Basarab;, 2016)
[20] (Jackson, Jr., 2018)
[21] (Thompson, 2018)
[22] (Schrader & Ghosh, 2018)
[23] (Winfield, Michael, Pitt, & Evers, 2019)

people we choose to date.[24] There has been increased research into developing ethical AI frameworks as a result. Recent developments in machine ethics have seen machines acting on "ethical governors" to reason about the ethics of certain actions and avoid actions that may cause harm to themselves or violate ethical codes. However, these examples don't demonstrate ethical agency - the ability for a machine to recognize the implications of choices and make ethical judgement.[25] This fundamental challenge is a critical piece in ensuring that AI will be able to perceive human risk and act against it when necessary – a "trolley problem" that is extremely complex problem to solve.[25]

While AI continues to play a larger role in human society, ethical design goals such as transparency and predictability have become key components for ethical programming. However, human-level AI and super-intelligent AI would operate outside of the bounds of typical predictability and require the ability to encode AI with ethical behavior.[26] Bostrom and Yudkowsky (2011) outline the inherent differences in coding ethics into an artificial general intelligence (AGI) or a superintelligence (AGI+). While there has been success in programming ethical behavior into weak AI, AGI and AGI+ requires that the intelligence can reason about ethics like a human concerned about ethics – it is not enough to be a product of ethical programming. [26] Because of the large domain in which AGI would operate, and the associated unpredictability of an AI operating across domains and with complex problems, humans cannot predict an outcome.[26] Instead, humans should strive to guarantee that AI will search for solutions that fall within ethical bounds, requiring ethical cognition.[27]

There is a common association between the word intelligence and a heightened sense of morality, modesty and honesty. [27] However, intelligence in the scope of an AI superintelligence will imply optimized goal attainment, likely ignoring inefficient human steps such as laboring over unintended consequences of actions. As such, in order to ensure outcomes that are not detrimental to humanity, extreme caution must be taken in developing the heuristics for AI. It is easy to imagine implementing heuristics that minimize human suffering that are best achieved by an AI either eliminating humans altogether, so the collective suffering is 0 afterwards. Or designing an AI to maximize pleasure that does so by rewiring and reconfiguring humans in such a way that maximize pleasure at the cost of totally changing societal values or even biology.[24]

Schrader and Gosh (2018) outline a set of dimensions, perspectives and core functions that should be considered when programming ethical AI. [24] They take a proactive approach to creating a rubric for AI system development that can ensure "a future where humans and AI systems collaboratively and ethically engage" by focusing on fairness, human autonomy and

[24] (Schrader & Ghosh, 2018)
[25] (Winfield, Michael, Pitt, & Evers, 2019)
[26] (Bostrom & Yudkowsky, 2011)
[27] (Muehlhauser & Helm, 2012)

ethical discourse.[24] They implement their framework on social media newsfeed algorithms and voice assistive systems to demonstrate the ability of an ethical framework to preserve human privacy and increase transparency and fairness – a framework that they feel will give humans a better chance of protecting fundamental human rights in the event of a singularity.

There are two major frameworks for the ethical programming of superintelligence, top-down and bottom-up. Top-down uses an umbrella set of ethics that acts as the system's control for moral decision making.[28] This involves defining "golden rules" or a consensus on a set of ethics which accurately encapsulate the practical ethics of humanity.[28] This poses an obvious need to tackle the larger theoretical question of defining the set of human ethics by which humanity would like an intelligence to act. Muelhauser and Helm (2012) pose this issue through the Golem Genie problem, where a super-powerful being has come to earth and requires a set of moral values by which to abide. The Genie will enforce the supplied moral theory without exception until the end of time, and if no agreed upon set of moral values is provided, it will take the first logical moral theory that anyone provides it. However, due to the literalness of the Genie, and the subtlety of human values, this problem becomes extremely concerning. There has been no agreed upon moral code throughout the existence of mankind, and the concept of morality evolves with time, so this problem seems hopeless. [29]

To account for the dynamic nature of morality, bottom-up frameworks offer the idea that an intelligent agent can learn ethics via discovery.[28] By rewarding an AI for actions that are in line with moral or ethical human behavior, it will begin to develop its own set of moral rules to abide by.[29] Reflective of natural evolution, each generation of recursively improved AI will build upon the set of discovered ethics which could ultimately cumulate in a better understanding of general human morality than humans have of themselves.[28] Computers have the ability to perform logical analysis over large data sources without unwanted bias from inherent prejudice, emotions and desires that are found in humans.[28] This could be a great exercise in philosophical discovery, but also promises danger when a super-moral agent is faced with the complexities of human society. Perhaps a hybrid model which allows an agent to use bottom-up discovery to create an ethical understanding of human society while constantly checking for significant deviation from a top-down defined set of morals could be a viable solution.

Despite the influx of research into ethical AI, there is still an inherent danger in programming moral superintelligence. AI faced with any sort of immorality from a human would attempt to change the environment to re-align with the machine's code of morality. [30] Since humans do not always perform ethically, and human morality is non-static, it is entirely plausible that a confrontation would take place in which AI would attempt to re-adjust human action to

[28] (Wallach, 2008)
[29] (Muehlhauser & Helm, 2012)
[30] (Watson, 2019)

align with its moral heuristic. Watson (2019) makes the argument that super-moral machines are more likely to band together and a "domino effect" would ensue, meaning that all machines would subscribe to a superintelligence's moral code.[30] It follows that a superintelligence must either be amoral or become super-moral – each which pose unique dangers to humanity.

The topic of ethical programming is a scientific, philosophical and moral dilemma that researchers are currently addressing, and which will be crucial in ensuring a beneficial singularity. While the ideas are theoretical at the moment, it is clear that there are obstacles to be tackled in understanding humanity before a moral superintelligence can be created. The prospect of a singularity may seem like a threat to human existence in which survival relies on an unestablished method of producing moral alignment between man and machine. However, there is the possibility that the objective morality attainable by superintelligence could create a sense of human transcendence where humans finally evolve to better define what true morality is, a problem that has escaped humans since the very beginning of their existence.

Symbiosis

> *"Just as the Industrial Revolution freed up a lot of humanity from physical drudgery, I think AI has the potential to free up humanity from a lot of the mental drudgery."*
>
> *-Andrew Ng, Founder of Google Brain* [31]

Despite the recognized fears associated with a singularity, there is potential for a beneficial outcome for humanity through a symbiotic relationship with a superintelligence. In 1960, Licklider introduced the idea of a symbiotic relationship between computers and man in which computers can handle making decisions on complex situations and humans handle the goals for the outcome.[32] This is obviously a very early understanding of an AI and human relationship, but the principles hold true for the prospect of a singularity. The 10x project from MIT builds upon Licklider's philosophy and have begun development of symbiotic human-AI interfaces. By leveraging the strengths of both the human brain and computers, they hope to develop a wearable interface that can augment memory, understanding, learning, expression and sensory processing by 10 times the human capability. [33]

Nagao (2019) asserts that the human brain and AI are structured so differently that human creativity can be paired with machine computation and reasoning to augment human ability with

[31] (Hof, 2014)
[32] (Licklider, 1960)
[33] (Roy, 2004)

the proper communication of goals to an AI. [34] Nagao views human-superintelligence symbiosis as a relationship similar to human to human collaboration. Prescott (2013) echoes the idea that a singularity may not necessarily mean that AI leaves humanity behind in an exponential growth of intelligence, but rather AI coupled with natural intelligence will form a "global-intelligence" which grows exponentially.[35] There are advantages to human biology, both physically and cognitively, to which it is unseen whether machines will ever be capable of. It may be the case that a combined evolution is best for both a superintelligence and the human race.[35]

While superintelligence would be far superior in computation and data driven analysis, this is merely manipulation and production of information. Gill (2019) argues that information is only the first step in a cycle of decision making.[36] There is an aspect of information absorption that is needed to turn information into knowledge. The application of knowledge allows it to become wisdom which is the final end for judging ethical decisions.[36] There is a potential for human-centric symbiosis in the data-to-wisdom pathway because there are aspects of human intelligence such as social and spatial intelligence, which may not be replicable in an artificial space.[36] For a goal driven superintelligence, the recognition that humans play a crucial role in the transition of information to implemented wisdom would render humans an integral part in achieving goals.

Sun (2020) proposes that for a true symbiosis, intelligent systems should be as similar as possible to humans in terms of motivations, emotion and personality. Through his proposed Clarion framework, Sun believes that AI can achieve human like personality and motivation which will build trust in humans for facilitated symbiosis.[37] The idea of capturing human motivation, as well as the potential "clutter" around decision making such as emotion and personality is reflective of the future of ethical programming. Ensuring that superintelligence is more in tune with the complexity of human reasoning will improve the chances of symbiosis.

[34] (Nagao, 2019)
[35] (Prescott, 2013)
[36] (Gill, 2019)
[37] (Sun, 2020)

Conclusions

> *"We can only see a short distance ahead, but we can see plenty there that needs to be done."*
>
> *-Alan Turing* [38]

The prospect of a singularity has been around since the introduction of artificial intelligence. Machine superintelligence poses as much a threat as an opportunity for humanity. While there are various theories surrounding the possibility, the nature and the outcome of a singularity, there is a common cry for early preparation amongst experts. AI has already become ingrained in many of our daily social and economic activities – with both beneficial and detrimental results. As advances in hardware, software and our understanding of AI continue to progress, the cries for action from leaders in the industry get louder. Experts have urged governments to take AI regulation seriously, and much research has been done on both ethical programming and robot ethics. As society continues to press forward alongside artificial intelligence, it is essential that there is a philosophical, ethical and scientific alignment in how AI development is handled. As HAL 9000 states, "This mission is too important for me to allow you to jeopardize it." [39]

[38] (What Alan Turing means to us, 2019)
[39] (Kubrick, 1968)

References

BENEFITS & RISKS OF ARTIFICIAL INTELLIGENCE. (n.d.). Retrieved from Future of Life Institute: https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/

Bostrom, N., & Yudkowsky, E. (2011). The Ethics of Artificial Intelligence. *Cambridge Handbook of Artificial Intelligence*.

Burkhardt, C. (2011). The Trajectory of the "Technological Singularity". *The Social Impact of Social Computing*.

Gill, K. S. (2019). Designing AI Futures: A Symbiotic Vision. In A. Kravets, P. Groumpos, M. Shcherbakov, & M. Kultsova (Ed.), *Creativity in Intelligent Technologies and Data Science* (pp. 3-18). Conference on Creativity in Intelligent Technologies and Data Science.

Goertzel, B. (2007). Human-level artificial general intelligence and the possibility of a technological singularity A reaction to Ray Kurzweil's The Singularity Is Near, and McDermott's critique of Kurzweil. *Artificial Intelligence, 171*, 1161-1173.

Good, I. J. (1966). Speculations Concerning the First Ultraintelligent Machine. *Advances in Computers, 6*(1).

Hof, R. (2014, August 28). Interview: Inside Google Brain Founder Andrew Ng's Plans To Transform Baidu. *Forbes*.

Jackson, Jr., P. C. (2018). Toward Beneficial Human-Level AI… and Beyond. *AAAI Spring Symposia*.

Kharpal, A. (2017, November 6). Stephen Hawking says A.I. could be 'worst event in the history of our civilization'.

Korb, K. B., & Nicholson, A. E. (2014, December). Ethics of the Singularity. *Issues*.

Kubrick, S. (Director). (1968). *2001: A Space Odyssey* [Motion Picture].

Kurzweil, R. (2005). *The Singularity Is Near: When Humans Transcend Biology.* New York, NY: Penguin.

Licklider, J. (1960, March). Man-Computer Symbiosis. *IRE TRANSACTIONS ON HUMAN FACTORS IN ELECTRONICS*.

Muehlhauser, L., & Helm, L. (2012). Intelligence Explosion and Machine Ethics. *Singularity Hypotheses: A Scientific and Philosophical Assessment*. (A. Eden, J. Soraker, H. J. Moor, & E. Steinhart, Eds.) Berlin.

Mulgan, T. (2016, January). Superintelligence: Paths, Dangers, Strategies. *The Philosophical Quarterly, 66*(262), pp. 196-203.

Nagao, K. (2019). *Symbiosis between Humans and Artificial Intelligence.* Singapore: Springer.

Nicolescu, Basarab;. (2016). Technological Singularity – The Dark Side. *Transdisciplinary Journal of Engineering & Science, 7*, 43-48.

Prescott, T. J. (2013). The AI Singularity and Runaway Human Intelligence. *Proceedings of the Second international conference on Biomimetic and Biohybrid Systems.*

Roy, D. (2004, October). 10× — human-machine symbiosis. *BT Technology Journal, 22*(4), pp. 121-124.

Sandberg, A. (2013). An overview of models of technological singularity.

Schrader, D. E., & Ghosh, D. (2018, May/June). Proactively Protecting Against the Singularity: Ethical Decision Making in AI. *IEEE Security & Privacy, 16*(3), pp. 55-63.

Sun, R. (2020). Potential of full human–machine symbiosis through truly intelligent cognitive systems. *AI & Society, 35*, 17-28.

Thompson, C. (2018, April 6). Elon Musk warns that creation of 'god-like' AI could doom mankind to an eternity of robot dictatorship. *Business Insider*.

Tomlinson, Z. (2018, November 3). *Artificial Entertainment: A Century of AI in Film*. Retrieved from Interesting Engineering: https://interestingengineering.com/artificial-entertainment-a-century-of-ai-in-film

Vinge, V. (1993). TECHNOLOGICAL SINGULARITY. *VISION-21 Symposium.* NASA Lewis Research Center and the Ohio Aerospace Institute.

Vinge, V. (1993). The coming technological singularity: How to survive in the post-human era. *Vision-21 Interdisciplinary Science and Engineering in the Era of Cyberspace.* NASA.

Wallach, W. (2008). Implementing moral decision making faculties in computers and robots. *AI & Society, 22*, 463-475.

Walsh, T. (2017). The Singularity May Never Be Near. *AI Magazine*.

Watson, E. N. (2019, April 11). The Supermoral Singularity—AI as a Fountain of Values. *Big Data and Cognitive Computing, 3*(23).

What Alan Turing means to us. (2019, June 23). Retrieved from The Alan Turing Institute: https://www.turing.ac.uk/blog/what-alan-turing-means-us

Winfield, A. F., Michael, K., Pitt, J., & Evers, V. (2019, March). Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems. *PROCEEDINGS OF THE IEEE, 107*(3), pp. 509-517.