# Pacing Progress; Mitigating Risks in the Rapid Evolution of AI/ML Systems

Andrew E. Lawrence
Tufts University
ENG-0001-22

November 7th, 2022

*Addendum:*
*This essay was originally written for my first-year writing seminar, ENG-0001-22, with Pelin Kivrak. It has since been edited for clarity, however, the tone and message remain the same.*

*Note that I use AI & ML interchangeably throughout this essay. Artificial Intelligence is an umbrella term which covers and intersects with many fields, including Machine Learning.*

## Introduction

In "The Sentient Machine", Amir Husain argues that OpenAI initially popularized the movement for transparent artificial intelligence (AI) development by making their "designs and code publicly available"[1]. Yet in recent years their organization has largely been viewed as the exemplar of skewed AI initiatives—where an initially powerful idea, transparent development, is brought down by foundational challenges to alignment and quick progress in the hopes of investment[2]. Because initial developments in machine learning (ML) were limited, such recent rapid advancements of agent capabilities have brought into light numerous previously unresolved hazards. Due to the outstanding challenges of establishing value-aligned AI systems, the development of broad use-case agents must be more thoroughly reviewed as it poses an enormous risk to the vitality of the technology industry and society as a whole.

## The Ethical Imperative

While the development of a universal moral compass for artificial intelligence is nonessential to the success of simpler artificial narrow intelligence (ANI) programs, as systems increase in complexity and scope, so does the necessity for value-alignment[3]. As a result, Crawford's conclusion that software developers' implicit biases greatly shape AI systems is incredibly problematic as AI's impact grows[4]. Such biases in development can cause serious consequences. For example, while ML detection systems, like those used in autonomous vehicles, are not yet

[1]Husain, Amir, *The Sentient Machine: The Coming Age of Artificial Intelligence*, (Scribner Book Company, 2018.), 18.
[2]Hao, Karen, *The Messy, Secretive Reality behind OpenAI's Bid to Save the World.* (MIT Technology Review, 2020.)
[3]Amodei, Dario, and Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, Dan Mané, *Concrete Problems in AI Safety.* (arXiv preprint, June 21, 2016.), 3, 21; Gabriel, Iason, and Vafa Ghazavi, "The Challenge of Value Alignment: From Fairer Algorithms to AI Safety", in *The Oxford Handbook of Digital Ethics*, ed. Carissa Véliz, (Oxford Academic, March 18, 2022.), 1.
[4]Crawford, Kate, *Atlas of AI: Power Politics, and the Planetary Costs of Artificial Intelligence,* (Yale University Press, 2021.), 13.

commonplace outside of testing and military applications, studies show they are notably less accurate in identifying participants with darker skin tones due to inequitable representation in test groups[5]. This can result in disastrous situations if broadly applied, and therefore must be addressed urgently. In "The Challenge of Value Alignment: from Fairer Algorithms to AI Safety," Gabriel et al. take Crawford's conclusion a step further, implying that because values are "embedded in technologies" it is imperative that engineers prioritize "value sensitive design": the practice of accumulating the needs of many communities[6]. While the assessment of Crawford and Gabriel et al. that artificial intelligence is influenced by the dominant interests of its designers may be accurate, there are those that push back, instead suggesting that complex AI will naturally develop "proper" beliefs[7]. However, Nick Bostrom, a leading mind in AI development, indicates otherwise. Rather, his widely accepted orthogonality thesis would suggest that there is no correlation between intelligence and alignment[8]. This means that as ANI systems approach the artificial general intelligence (AGI) classification, no one should expect the alignment issue to disappear. As a result, the need for an installable ethical framework, one that would drive AGI systems in the "right" direction, is increasingly apparent and necessary as these agents grow in influence.

Challenges in Building Ethics Frameworks

While the necessity for an ethical template for agents to follow has been established, it is easier said than done; the development of any structure, either through proper representation or dictatorial decision-making would be a serious challenge. When discussing the overarching issues with developing safe ML systems, Aliman et al. present the main issues with the creation of a comprehensive model of humanity's values. In order to develop a genuine framework an extensive "ethical self-assessment [and] debiasing" would need to be done[9]. However, Aliman et al. believe the clarification of human values would be a challenge since "humans are often reluctant to clearly express what they want."[10] Although this analysis is more speculative than concrete, they do illustrate the crux of the issue: developing such a comprehensive model would require the coordination and collaboration of millions, if not billions. This achievement is highly improbable considering local governments may spend years bickering over the correct placement of store signs, let alone meta-ethics. As a result, many developers attempt to circumvent this by trying to implement broad decision-making theories. Of those, augmented utilitarianism, an ethical structure based on its namesake, has proven popular. Its success comes from its ability to limit the perversion of a programmer's original intentions by extending beyond

[5]Aliman, Nadisha-Marie, and Leon Kester, Peter Werkhoven, Soenke Ziesche, "Sustainable AI Safety?" In *Delphi - Interdisciplinary Review of Emerging Technologies*, vol 2, no. 4, 2019, eds. Ciano Aydin, Clara Jausin, and Jakob McKernan (Lexxion, April, 2019), 231.
[6]Gabriel, "The Challenge of Value Alignment", 3.
[7]Ibid., 11.
[8]Ibid.
[9]Aliman, "Sustainable AI Safety?", 229.
[10]Ibid., 227.

limitations of "consequentialist and utilitarian utility functions"[11]. While this flexible scaffold seemingly contains popular moral decisions, Aliman et al. go on to state that augmented utilitarianism is limited by many "sensor-level issues"[12]. Given these difficulties, either approach to establishing a moral framework for agents seems unlikely without an industry-wide consensus and systematic analysis of the fundamental application of ethics in AI systems.

Even if the world successfully agreed on a correct set of values, their implementation would present another challenge as the two most common methodologies of ML alignment, the top-down and bottom-up approaches, are riddled with difficulties. The top-down procedure is generally associated with a clear moral theory that is then implemented in an algorithm[13]. Yet, not only does the top-down approach still have to contend with the realization of a dominant ethical framework, but in its application "moral rules" often come into intractable conflict with other rules[14]. Furthermore, Gabriel et al. contend that if the "ultimate goal is social value alignment," then this method is problematic because it contains an inherent bias towards the most compatible, and therefore not representative, option[15]. Given the various difficulties with the top-down approach, other engineers have looked to lead agents to value alignment by reinforcing "praiseworthy conduct"[16]. However, this bottom-up method does not have specific goals or rewards, resulting in the development of a black box of data—where researchers can only see the initial and final points of a decision[17]. Not only does this become a serious problem during the application of any bottom-up algorithm as developers are unable to extrapolate information from the data, but also it represents a legal issue as companies cannot explain to their users why a choice was made. As a result, engineers should focus their efforts on resolving the technical aspects of the alignment issue as current methodologies do not legitimately develop transparent and properly aligned agents.

Reward Hacking and Distributional Shifts

While the issue of misaligned ML systems illustrates a dangerous aspect of modern agents, the increasing influence of agents constitutes a need to mitigate side effects created through reward hacking and safe exploration, among others. Although an agent may technically succeed in fulfilling their function when caught reward hacking, it is an important issue because it subverts the designer's overarching intent by shortcutting the program[18]. Consequently, solving the fundamental issue behind reward hacking is an important step to take to ensure safe application. To this point, the fact that decoding reward hacking appears to be a serious challenge as it is tied to the

[11]Aliman, "Sustainable AI Safety?", 228; See Aliman, Nadisha-Marie, and Leon Kester, "Augmented Utilitarianism for AGI Safety." In *Artificial General Intelligence: 12th International Conference, AGI 2019, Shenzhen, China, August 6-9, 2019, Proceedings,* eds. Patrick Hammer, Pulin Agrawal, Ben Goertzel, and Matthew Ikle (Springer Cham, July 25, 2019), 1.
[12]Aliman, "Sustainable AI Safety?", 229.
[13]Gabriel, "The Challenge of Value Alignment", 8.
[14]Ibid., 9.
[15]Ibid.
[16]Ibid., 8.
[17]Ibid., 9.
[18]Amodei, *Concrete Problems in AI Safety,* 2.

basic principles of ML is especially concerning[19]. Furthermore, the necessity to prioritize accident reduction research is also demonstrated by the danger agents pose when undergoing any distributional shift. At the moment, exploration policies assume heuristics will apply to new situations and make "no attempt to avoid dangerous situations" when choosing actions[20]. While these characteristics may be acceptable in focused systems where change is unlikely, the growing power of agents to alter our way of life places enormous importance on resolving this issue to create safe systems. Yet the two approaches to resolving the issue, "covariate shift assumption" and generative distribution models, are both particularly fragile[21]. Therefore, developers should not assume that they have accurately reduced the risks involved with exploring new environments. This suggests that challenges of reward hacking and distributional shift in ML systems still represent a serious risk that engineers take on when implementing any AI.

Counterclaim

Although there is a consensus among researchers that the development of AI is too fast, there are those who believe that the current pace is required as the long-term benefits for humanity are overwhelmingly positive. Amir Husain's "The Sentient Machine" does acknowledge the risks posed by unaligned agents, yet, he strongly believes that significantly regulating AI work will be "incredibly harmful to us as a civilization."[22] He goes on to suggest it is "imperative we embrace" ML systems in all sectors as the comparatively infinite productivity of AI is essential to upgrading our standard of living[23]. Although the implementation of such systems will likely benefit humanity if done properly, at the moment such systems are woefully unprepared for widespread application. Not only do agents contain various unresolved alignment and accident management issues, such as those mentioned above, but there is also a widespread mistrust of artificial intelligence as a whole[24]. With these factors in mind, the sustainable integration of AI into different industries in the near future, and the subsequent race to achieve the quickest results, should be reconsidered.

Conclusion

From reward hacking to Augmented Utilitarianism, developing safe AI models is a controversial and complex issue. Yet the risk created by misaligned or accident-prone agents is undoubtedly impactful and calls for continuous reexaminations of the methods by which ML systems come to life. While, admittedly, the development of safe AI is an evolving field, one with both rapid successes and failures, the issues presented in this paper represent a small scope of those challenging researchers, and, there is still a noticeable lack of institution-wide consensus on any fundamental progress being made. If AI advocates want to push their agents onto the people of the world, they should do so by demonstrating their safety rather than ingenuity.

[19]Ibid., 8.
[20]Ibid., 14,16.
[21]Ibid., 17.
[22]Husain, *The Sentient Machine*, 18.
[23]Ibid., 13.
[24]Aliman, "Sustainable AI Safety?", 232.

# Bibliography

Aliman, Nadisha-Marie, and Leon Kester, "Augmented Utilitarianism for AGI Safety." In *Artificial General Intelligence: 12th International Conference, AGI 2019, Shenzhen, China, August 6-9, 2019, Proceedings,* eds. Patrick Hammer, Pulin Agrawal, Ben Goertzel, and Matthew Ikle (Springer Cham, July 25, 2019). https://doi.org/10.1007/978-3-030-27005-6_2

Aliman, Nadisha-Marie, and Leon Kester, Peter Werkhoven, Soenke Ziesche, "Sustainable AI Safety?" In *Delphi - Interdisciplinary Review of Emerging Technologies*, vol 2, no. 4, 2019, eds. Ciano Aydin, Clara Jausin, and Jakob McKernan (Lexxion, April, 2019). https://delphi.lexxion.eu/article/DELPHI/2019/4/12

Amodei, Dario, and Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, Dan Mané, *Concrete Problems in AI Safety.* (arXiv preprint, June 21, 2016.) https://doi.org/10.48550/arXiv.1606.06565

Crawford, Kate, *Atlas of AI: Power Politics, and the Planetary Costs of Artificial Intelligence,* (Yale University Press, 2021.)

Gabriel, Iason, and Vafa Ghazavi, "The Challenge of Value Alignment: From Fairer Algorithms to AI Safety", in *The Oxford Handbook of Digital Ethics*, ed. Carissa Véliz, (Oxford Academic, March 18, 2022.) https://doi.org/10.1093/oxfordhb/9780198857815.013.18

Hao, Karen, *The Messy, Secretive Reality behind OpenAI's Bid to Save the World.* (MIT Technology Review, 2020.) https://www.technologyreview.com/2020/02/17/844721/ai-openai-moonshot-elon-musk-sam-altman-greg-brockman-messy-secretive-reality/.

Husain, Amir, *The Sentient Machine: The Coming Age of Artificial Intelligence*, (Scribner Book Company, 2018.)