Presidential Duties in Addressing the Advent of Artificial General Intelligence

Humanity is on the advent of creating artificial general intelligence (AGI), a form of AI that can meet or exceed human performance in any task (Ord 143). With experts in the field predicting that AGI will be deployed within the coming few decades, a prudent approach must be taken toward responsibly developing and regulating this technology, as AGI presents an existential risk to humanity. Effectively addressing the risks that stem from AGI will necessitate a wide number of resources and coordination of individuals from varied disciplinary backgrounds. Thus, this essay will address the risk from the position of a sitting US president as the technology is being developed.

To develop an approach that addresses the existential risk that AGI presents, it is necessary to first discuss the nuances that make AGI an existential risk. This can be separated into two main issues: current incapacities to align AGI to human values, and the integration of AGI into military applications.

Regarding the alignment issue, the inability to instill human values onto AGI is incredibly dangerous as it leaves open the capacity for AGI to determine that humanity is inconsistent with the goals of that AGI, thus motivating it to take control of humanity's future for its own ends (Ord 145). This is a problem due to the reward function that is used to train AI, as it does not contain any easy means of instilling human values. Thus, the result would be unaligned AGI that operates via a purely consequentialist framework, serving only to maximize its reward function without any means of utilizing Aristotle's Golden Mean to determine and uphold the virtuous behavior that Ord argues is necessary to preserve and protect humanity's potential (Ord 53). To uphold virtuous behavior, the Golden Mean requires being able to strike a balance between the

excesses and deficiencies in character; however, AGI that can only work toward maximizing a certain outcome will be unable to strike this necessary balance (Maden).

The concerns associated with unaligned AGI are compounded due to the inevitable adoption of AGI into military usage, a circumstance under which the actions of unaligned AGI may be even more detrimental to humanity due to its inability to uphold human values enshrined in international humanitarian law (Maxwell, Walsh). Furthermore, AGI in warfare scenarios would dramatically increase the chances of an unplanned conflict occurring. This means that the existential risk factor of great-power war, which Ord considers as being "a larger contributor toward existential risk than most of the specific risks" that he examines, would thereby increase as well (Ord 176). The result is that unaligned AGI is not only a direct existential risk, but also serves to perpetuate other risks through increasing the existential risk factor of global-power war. Therefore, it is essential to address this issue.

Before introducing the approach that the sitting president ought to take toward addressing the existential risk that AGI poses, an important consideration must be addressed: halting the development of AGI entirely is not a practical solution. Despite the existential risk, it is essential for humanity to develop AGI due to the transformative effect that it could have on the world, through developing more efficient solutions to world problems, as well as identifying and mitigating other risks (Ord 148). In fact, failure to develop AGI due to fears of the existential risk it presents would be a violation of teleological ethics, as it would prevent humanity from reaching its full potential, contradicting the very purpose of preventing existential catastrophe in the first place – to preserve humanity's full potential (The Ethics Centre, Ord 6). Thus, attempting to halt the development of AGI would be both an unethical and unpragmatic approach

in this circumstance. With this consideration accounted for, the optimal approach can be presented.

Ord emphasizes the cruciality of international coordination in resolving issues of existential risk (Ord 199). Therefore, to achieve this end, the president ought to advocate for the creation of a new international institution centered around global cooperation on AGI. This new institution would coordinate efforts to develop international standards for AGI, consisting of carefully established norms and regulations collectively developed by experts in the field (Ramamoorthy 6). The institution also ought to be responsible for monitoring the development of AGI by both states and private entities to ensure compliance with the newly established standards. Additionally, with the bully pulpit, the president would be able to establish the importance of research on AGI alignment, and the institution can spearhead efforts to develop methods of aligning AGI (Merriam-Webster).

Central to this international approach would be the creation of an AGI through a collective international effort via this new institution, similar to international cooperation on the International Space Station (Ramamoorthy 5). This AGI would be developed with prudence and patience to ensure proper alignment with human values (after the proper means of achieving this had been developed), as well as accordance with the new international standards for AGI. This AGI developed through global cooperation can serve to improve global welfare and be dedicated toward mitigating other existential risks (Ramamoorthy 5). Additionally, this AGI can serve a regulatory role in ensuring that independently developed AGI conforms to international standards, resolving the existential risk that AGI used for military applications in specific states presents, as there would be a stronger international AGI that could counteract attempts to end humanity's potential (Haney 168).

Furthermore, the president ought to consider the two main formulations of Kant's Categorical Imperative when developing the institution and encourage engineers to instill these principles into the decision-making process of the international AGI. These two main formulations are the universality principle, and the respect for persons principle (Johnson). Fundamentally ingraining these principles into the AGI would ensure that its actions would truly be for the benefit of the international community, as well as being fundamentally aligned with the wellbeing of humans.

Inevitably, there are risks that would threaten this approach. For instance, rogue actors such as maligned individuals that sought to develop intentionally harmful AGI could potentially undermine the efforts of the new international organization (Ramamoorthy 4). However, a coordinated international effort into developing a powerful, regulatory international AGI would be able to counteract this issue. A graver risk would be a lack of international support for such an organization, preventing it from being created in the first place. This would mean that the existential risk presented by AGI would likely go largely unaddressed, which would be detrimental to humanity's future. A solution to this would be for the president to frame the development of the organization as an essential component for collective security in an AI-driven future, being necessary to avoid another era of mutually assured destruction, which entails a largely unstable peace (Ramamoorthy 2). This would likely prompt many nations to support the establishment of this organization, uniting the nations of the world under the banner of humanity against the existential threat that AGI presents if not addressed with prudence. While developing an international institution to respond to this risk is no easy task, it has been done before by past presidents. Therefore, an effort ought to be made for the future of humanity.

Works Cited

"Bully Pulpit." *Merriam-Webster*, 24 Apr. 2023, https://www.merriam-webster.com/dictionary/bully%20pulpit.

"Collective Security Can Be Assured by Finding Common Ground amid Primacy of National Positions, Members Told at Start of First Committee Session | UN Press." *United Nations*, 7 Oct. 2014, press.un.org/en/2014/gadis3497.doc.htm.

Compagnoni, Ares Simone Monzio. "Will Artificial General Intelligence Change the Nature of War?" *Military Strategy Magazine*, 2023, https://www.militarystrategymagazine.com/article/will-artificial-general-intelligence-change-the-nature-of-war/.

The Ethics Centre. "Ethics Explainer: Teleology." *THE ETHICS CENTRE*, 4 Apr. 2022, ethics.org.au/teleology/.

Haney, Brian Seamus. "The Perils & Promises of Artificial General Intelligence." *SSRN Electronic Journal*, 2018, https://doi.org/10.2139/ssrn.3261254.

Johnson, Robert, and Adam Cureton. "Kant's Moral Philosophy." *Stanford Encyclopedia of Philosophy*, 21 Jan. 2022, plato.stanford.edu/entries/kant-moral/#CatHypImp.

King, Anthony. "AI at War." *War on the Rocks*, 27 Apr. 2023, warontherocks.com/2023/04/ai-at-war/.

Maden, Jack. “The ‘Golden Mean’: Aristotle’s Guide to Living Excellently.” *Philosophy Break*, Jan. 2023, philosophybreak.com/articles/the-golden-mean-aristotle-guide-to-living-excellently/.

Maxwell, Paul. “Artificial Intelligence Is the Future of Warfare (Just Not in the Way You Think).” *Modern War Institute*, 20 Apr. 2020, mwi.usma.edu/artificial-intelligence-future-warfare-just-not-way-think/.

Ord, Toby. *The Precipice*. Hachette Books, 2020.

Ramamoorthy, Anand, and Roman Yampolskiy. “BEYOND MAD?: THE RACE FOR ARTIFICIAL GENERAL INTELLIGENCE.” *ICT Discoveries*, http://handle.itu.int/11.1002/pub/812a0228-en.

“The U.S. Constitution: Preamble.” *United States Courts*, www.uscourts.gov/about-federal-courts/educational-resources/about-educational-outreach/activity-resources/us#:~:text=%22We%20the%20People%20of%20the,for%20the%20United%20States%20of. Accessed 14 June 2023.

Wallach, Wendell. “The Techno-Military-Industrial-Academic Complex.” *Fortune*, 18 Feb. 2022, fortune.com/2022/02/18/techno-military-industrial-academic-complex-eric-schmidt-google-yale-defense-politics-universities-wendell-wallach/.

Walsh, Toby. “The Problem with Artificial (General) Intelligence in Warfare.” *Centre for International Governance Innovation*, 28 Nov. 2022, www.cigionline.org/articles/the-problem-with-artificial-general-intelligence-in-warfare/.

Walsh, Toby. “The Problem with Artificial (General) Intelligence in Warfare.” *Centre for International Governance Innovation*, 28 Nov. 2022, www.cigionline.org/articles/the-problem-with-artificial-general-intelligence-in-warfare/.

“What Is Effective Altruism?” *Effective Altruism*, www.effectivealtruism.org/articles/introduction-to-effective-altruism#fn-15. Accessed 14 June 2023.