BEYOND MAD?: THE RACE FOR ARTIFICIAL GENERAL INTELLIGENCE

Anand Ramamoorthy[1], Roman Yampolskiy[2]
[1]Dept. of Anaesthesiology, UW-Madison, WI, USA
[2]Computer Engineering and Computer Science, University of Louisville, KY, USA

Abstract – Artificial intelligence research is a source of great technological advancement as well as ethical concern, as applied AI invades diverse aspects of human life. Yet true artificial general intelligence remains out of reach. Based on the history of deeply transformative technologies developed by multiple actors on the global stage and their consequences for global stability, we consider the possibility of artificial general intelligence arms races and propose solutions aimed at managing the development of such an intelligence without increasing the risks to global stability and humanity.

Keywords – AGI, AI safety, AI Treaty, arms races, global coordination, MAD

1. INTRODUCTION

Artificial intelligence (AI) research has a decades-long history. However, AI systems have come to occupy the public imagination now more than ever, thanks to some remarkable recent technical developments with compelling public demonstrations, as well as an unprecedented level of hype and concerns of existential risk.

Applied AI stands poised to revolutionize multiple industries and avenues of human life. AI systems are expected to replace human drivers in autonomous vehicles, human combatants in warfare, and the relevance of the human operator to many forms of professional activity will be called into question as these systems become more proficient and truly autonomous.

A WIRED magazine article in 2014 designated Go as the "ancient game that computers still cannot win", with some experts predicting that computers would only be able to win in a decade or more [1]. Well before the predicted time, Deep Mind's AlphaGo [2], attracted significant attention owing to its success against Lee Sedol, a 9th Dan in the game of Go [3], and the world's leading player Ke Jie [4].

A Genetic Fuzzy Tree method based AI, ALPHA, which is currently a simulation tool, has emerged as a likely candidate for future military applications involving unmanned aerial vehicles as well as mixed squadrons of manned and unmanned vehicles in aerial combat [5].

While such specialist AI systems exist, and are evidently becoming more and more capable, the most exciting (and according to AI safety researchers, the most risk-laden) goal of AI research has always been what can be termed Strong AI, that is, an artificial intelligence that is truly general and not constrained to a task-domain. Such an artificial general intelligence (AGI) does not exist in our world yet, to the best of our knowledge.

The possibility of such an AGI leading to so-called artificial superintelligences (ASI) via intelligence explosion (a scenario where the AGI recursively improves itself to the point of exceeding human-level intelligence to an unpredictable extent), has led scientists, philosophers and technologists to consider the existential risks (to humanity) posed by the development of AGI. Unlike technologists and techno-futurists, AI and machine-learning researchers are more conservative in their estimates as to when such a system (or multiple such systems) can be expected on the scene. It is to be noted that the machine learning/AI research community is mostly unsure as to whether AGI would come to be or result in ASI. There also exist optimistic views on AI development which provide a healthy counterbalance to the apocalyptic visions commonly encountered in the media [6].

That said, 48% of the expert respondents in a recent survey [7] did agree that "society should prioritize work on minimizing the potential risks of AI". A recent report from the U.S Government's Office of Science and Technology Policy examines the state of the art in AI technology and provides an overview of the benefits and risks of AI, concluding with policy recommendations for the administration [8]. AI and AI safety are predicted to be vital to both economic development and national security [9]. It stands to

reason that such considerations cut across borders, and are shared by nation-state stakeholders in the AI sector. AI arms races can happen in the context of "narrow" AI systems as well as AGI development. In the former instance, the most salient manifestation of such a race would be adversarial programs, administered by militaries around the world, to develop intelligent controllers and autonomous weapons systems. This possibility has attracted tremendous attention from AI safety experts as well as ethicists who rightly fear the ramifications of such systems being introduced into human armed conflict. The second type of AI arms race concerns multiparty, adversarial programs aimed at developing the world's first AGI.

Here we examine the prospect of AGI development in terms of escalation and arms races between diverse actors and stakeholders: state actors, non-state (corporate) actors, and rogue actors. We conclude by providing policy recommendations aimed at mitigating the risks identified.

2. ARMS RACES AND AGI: BEYOND MAD?

The success of The Manhattan Project and the deployment of nuclear weapons by the United States military in the Second World War led to a new kind of international conflict, a nuclear arms race, where powerful nations sought to acquire the same destructive capabilities as the U.S.A. This resulted in a world where an unstable peace is kept alive, informed to a significant extent, by the doctrine of mutually assured destruction (MAD) in addition to global non-proliferation efforts. A detailed discussion of the current understanding of MAD and the status of nuclear non-proliferation is beyond the scope of this paper. It suffices to note that examining the case of MAD in its original context provides insights that can be used to understand the role of disruptive technologies in international conflict (for case studies of such technologies see [9]). AGI, if and when it happens, may well be the final disruptive technological development engineered by humans. AGI represents a level of power that remains firmly in the realm of speculative fiction as on date. It stands to reason that if true AI were achievable, state actors would be invested in achieving this and with priority if possible. Such a quest for priority might have disastrous consequences due to corner-cutting when it comes to safety, and has been described as "racing to the precipice" [10]. An AI arms race is often spoken of in the context of the development of autonomous weapons systems which become increasingly sophisticated, changing the face of warfare. Were we to adopt Clausewitz's observation that "war is the continuation of politics and with other means" [11] and examine international conflict, it becomes obvious that the role of AI would extend well beyond, and emerge well before, armed conflict. A nation equipped with a fully general AI, would stand to benefit in the negotiation of conflict and agendas, regardless of means. If said AI were both general AND endowed with the ability to act in the world (i.e., not merely an Oracle-in-a-box as some have proposed, see [12] for an analysis of AI confinement), then, all arguments pertaining to the existential risk posed by AI would apply. Having AI systems autonomously determine the deployment of weapons in armed conflict is one major route to potential catastrophe, but we would like to emphasize that matters are likely to become fraught even before this development.

AGI development would push the global security strategy beyond what is currently in place. In the event of human control over the AGIs (which is a problem worth examining in its own right), MAD would not be sufficient to avert catastrophe. This would be because of the greater complexity associated with AGI and the capabilities such a system would present to human controllers, for instance, the AGI of a particularly belligerent state could calculate the optimal means to mortally destabilize the economy of a rival state (however powerful), or develop weaponized code disseminated globally to control, if not disrupt, vital systems such as power grids and communication networks. In other words, the cyber-warfare capabilities of an AGI-assisted nation-state would pose a serious threat to global stability and humanity. The current forms of narrow AI are capable of interfering with communication services. AI-enabled surveillance across communication networks is likely to become the norm. AI tools with the potential to perturb or alter the content of communications are already in development (see https://lyrebird.ai/ for an interesting example in the context of mimicking human speech; see also: https://lyrebird.ai/ethics/). An AGI with access to the Internet and communication networks in general would be able to, depending on its objectives (or of those who deploy it), selectively impede communication across a certain network/region, or fabricate misinformation to probe human responses if it develops the objective to understand social impact of communication networks. Much as these scenarios remind us of science fiction, it is worth noting that we encounter reports of computational propaganda or technology-assisted disinformation with increasing regularity. On a more optimistic note, an AGI that is constrained to cooperate with humans could help

envision more efficient use of resources for optimizing the communication networks we have available, or design altogether novel and better architectures.

As we discuss below, the potential development of AGI is unlikely to be an exclusively state-funded affair. Current breakthrough AI systems all appear to be products designed by technology companies.
Given the fact that technology giants such as Google, Facebook and others are beginning to open-source their machine-learning tools (e.g. TensorFlow), it is well within the realm of possibility that non-commercial, non-state actors including individuals would be able to develop applied artificial intelligence.

2.1. Actors in the AGI race

The development of AGI (or AGIs) could be due to the actions of many actors, and each type of genesis would likely be linked to the emergence of risks shaped by the intentions of the actors, the corresponding core objectives of the AGI and the context of AGI deployment. Here we examine possible scenarios as a function of actors and likelihood.

2.1.1. State actors

States function as agents/actors in the international environment, and pursue policies that benefit them as well as their allies whilst potentially mitigating the risks posed by states that are not allies or neutral, if not diminishing their influence regionally/ internationally. Global coordination remains a non-trivial challenge and one that is not easily resolved.
Imagine throwing in one state which suddenly had an AGI Oracle that could examine scenarios exhaustively for any given situation, access all relevant information from past situations of the sort being considered, and glean insights from actionable intel available to the highest executive in the state to guide decision making. If the Oracle has no influence on the world physically, and the state concerned is one with more to gain from a better solution to the global coordination problem, their use of the Oracle's counsel would likely benefit humanity.

It is conceivable that this may not be the most likely scenario. Powerful technology has often been a driving force in the pursuit of a world order where the nation-state with the technology in question is in a privileged position to pursue its objectives. If the state in question is less interested in achieving better global coordination or more interested in exercising disproportionate influence globally, the AGI would provide such a state actor with a potentially incalculable advantage.

Contemporary trends in various nations across the globe evidence a resurgence of nationalistic themes in politics, with elections delivering governance to those who promise a stronger nation "above all else". The world is not presently geared to support a single global community, and in a setup where nationalistic impulses influence both intranational politics and geopolitics, it is neither realistic nor prudent to assume that the achievement of AGI by any single state actor would be beneficial to humanity. It stands to reason that multiple state actors would seek unprecedented strategic advantage through AGI. If there is no commensurate development on the global coordination problem which renders the balance of power stable, this scenario would lead to catastrophe.

2.1.2. Corporate actors

Industry remains the face of AI research. With multiple corporate entities vying to develop true artificial intelligence, or artificial general intelligence, there already is a race to harness the power of AI for commercial ends whilst ostensibly impacting the world positively. Recent calls for research on AI safety have been made by researchers as well as leaders in the tech industry. Such a confluence of academic and corporate/industrial camps on existential risk posed by AI bodes well for research and development on this front. That said, the competitive and technological advantage presented by achieving priority in the development of AI is likely to incentivize some corporate actors to compromise on safety, resulting in unregulated, unsafe development of AI/AGI. One is reminded of the reported discrepancy between management and engineering divisions at NASA on the risk associated with continued operation of the space shuttle, as discussed in Appendix F of the report on the Challenger explosion [13]. Also, scientific and ethical consensus may not be sufficient to motivate technologically capable enterprises to focus on safety prior to developing AI across the globe. The problem of global coordination would remain a factor that would increase degrees of freedom in any given scenario where actors interact internationally.

2.1.3. Rogue actors

Unlike other dangerous technological developments in history, AI breakthroughs may not occur exclusively

in academic, governmental or industrial centers of research. In principle, powerful AI systems could be developed by individuals or groups with no national or corporate agendas. Such *homebrewed* AI would be hard to deal with precisely due to a lack of oversight, monitoring or consensus on architectures and objectives. More worrisome is the prospect of such rogue actors developing AI without safety considerations or with malicious intent (for example, see the case of commercially available unmanned aerial vehicles being repurposed for guerrilla warfare by terror groups [14]). It could be argued that the resources required to develop a powerful and truly general artificial intelligence may not be available to rogue actors, but it is far from clear that this would be the case, and it may be unwise to presume that any such obstacles would be insurmountable.

Cyberattacks originating from individuals/small groups are commonplace in our increasingly interconnected world, and it is conceivable that the development of an artificial general intelligence by rogue actors would be similar in terms of execution, but more harmful in terms of impact on human society and life. As a case in point, consider the recent large-scale spread of the WannaCry ransomware, exploiting a vulnerability in the Windows operating system (particularly versions past). Investigations seem to suggest that the architects of the attack were not well organized and the attack not as nightmarish as it could have been. Yet, it precipitated a significant amount of chaos and affected networked computers worldwide. Now replace the ransomware with an AGI that is released into the wild by hackers motivated by political ideology, notoriety or curiosity [15]. Even if the AGI is not inherently dangerous, the consequences in such a scenario would be hard to predict or plan for and could be catastrophic.

3. AGI AND VALUE ALIGNMENT

Researchers working on making any potential AGI "friendly" or compatible with human existence, if not values, speak of the AI value alignment problem. There are technical as well as pragmatic considerations attendant upon AGI research (see [16]), which increase the complexity of any proposed attempt to align the values (or objectives) of a putative AGI with human ones. The most obvious, and most non-trivial consideration (if not constraint), is the fact that humans across the world are not capable of value alignment to the extent that actions which increase existential risk for all humanity would be rendered extremely unlikely. Climate change policy is one example of a failure of global coordination on a matter of utmost importance. The current near-impossibility of universal nuclear disarmament is another. International conflict is never devoid of human cost. It appears to be the case that several developments in the AI sector are likely to exacerbate inequities in the bargaining power of nations on the international scene, as well as their ability to administer coercive force via an AI-augmented military. Given the lack of value alignment within human groups, it would be highly prudent to seek a solution to this problem in parallel with, if not prior to, working on AGI value alignment. It is to be noted that discussions on AGI emergence and global coordination posit the scenario of a singleton (one global governing entity), which appears markedly less likely to happen relative to the birthing of an AGI. In all likelihood, even the most benign AGI would be developed within a world where human groups (nation states) do not see eye-to-eye on several crucial issues. Given this, it would be wise to temper any optimism on the AGI front with a healthy appreciation for risks, safety concerns and the need to respect reality, particularly with regard to international conflict and human factors.

4. SHAPING AGI RESEARCH

[17] discusses the means and measures to shape AI (and potential AGI) research to promote safe and beneficial AI development, and makes the distinction between extrinsic measures such as constraints on design and intrinsic measures such as inducing a normative shift towards wanting to build beneficial AI and creating a stigma around dangerous AI research. [17] makes the compelling argument that hard, extrinsic measures such as outright bans might have effects counter to what was intended, (as a ban would draw attention to the problem in a manner that evokes curiosity and desire for boundary breaking). The intentional shaping of AGI research by targeting the culture in the research communities, both academic and non-academic, to make the wish to build safe and beneficial AI a social norm, with strong normative factors encouraging the avoidance of unsafe designs etc. appears to be an interesting strategy and one that is likelier to have a stronger impact over time, especially in the context of open AI development.

[18] discusses the implications of open AI development and elaborates on the complexities inherent to the pursuit of openness as a policy across multiple dimensions, such as the political, scientific and technological and it appears that the solution proposed here would make concerns regarding long-term costs of openness irrelevant, given that, assuming this solution is workable, a coalition of

states and researchers would work on AI (AGI) development as a public, open enterprise, inspired by ventures such as OpenAI.

[19] presents a formal account of determining a Pareto optimal policy by a machine built and deployed by two actors (individuals, companies or states), reflecting the beliefs and utility functions associated with each actor, and demonstrates that such a policy would sequentially shift the prioritization of one actor's utility function over the other's as a consequence of the accuracy of the actors' beliefs regarding the input to the machine (i.e. the state of the world in which the machine functions). This analysis raises the question whether the fruits of strategic cooperation in AI development could still be distributed unequally if the actors involved do not have access to either the same information regarding the machine's world (and inputs), or if one or more actors actively shape the beliefs of the other actors through disinformation.

5. PERSPECTIVES AND SOLUTIONS

In examining the potential for an AGI "arms race", and the trends observed in the present-day world, we see that there are multiple paths to averting such conflict in the decades to come. They are not mutually exclusive and can be integrated to form a comprehensive strategy. We recommend developing a solution that is layered, with failures at one level compensated for by policies and mechanisms in place at another, for redundancies and multiple defenses would render the solution robust.

5.1. Solution 1: Global collaboration on AGI development and safety

Postwar advancements in space science led to both the Space Age, characterized by the race for space. The erstwhile Soviet Union and the United States of America vied with each other for priority in space exploration and successful manned missions. While spacefaring nation states pursue space programs of their own with national agendas and strategic goals, several have also come together to create and sustain an international space station (ISS), which has become a new benchmark for international cooperation [20]. Taking this as a paradigmatic case of successful international cooperation, we refer to this as the ISS pathway.

If the ISS pathway is chosen with coordination and foresight, an AGI arms race could be avoided altogether by adopting the safe and beneficial development of AGI as a global, non-strategic humanitarian objective, under the aegis of a special agency within the United Nations, established for this exclusive purpose. In this scenario, countries would supply resources and invest equitably in the creation of the world's *first and only* AGI with safety considerations imposed at every level. As a first step towards securing such global cooperation, a comprehensive Benevolent AGI Treaty must be developed and ratified by all member nations of the UN. Recent calls for a ban on autonomous weapons are a step in this direction (see [21]). The 23 principles, enunciated in conjunction with the 2017 Asilomar Conference, provide a foundation for such a treaty (see: [22]). If such a treaty succeeds, any potential beneficial AGI would be treated as a global, public good with equal distribution of (carefully minimized) risks and (carefully maximized) benefits and no room for monopolies, adversarial co-opting of the system's potential, etc. The latter constraints are vital to the success of the treaty as an instrument and even more important to its efficacious implementation, as an AGI would represent a level of power hitherto unprecedented. Additional safeguards need to be researched on account of the fact that an AGI would be much easier to reproduce compared to other disruptive technologies humans have developed thus far. Further research needs to be done on how to convolve AI safety approaches such as "boxing in the AI", with collaborative AGI development, to prevent possible unauthorized, undesirable reproduction of the AGI by actors not sanctioned by the treaty.

The game theoretic analysis presented by [10] appears to support a cooperative approach, as teams working together had positive implications for safety, as opposed to teams racing to succeed. However, they also present an intriguing result that increased knowledge of the work being done by other teams increased risk. It is to be noted that this informational hazard would apply only if nations seek to gain strategic dominance or advantage, and a concerted, transparent effort to build an AGI for global welfare, is expected to improve safety and risk-mitigation efforts.

5.2. Solution 2: Global Task Force on AGI to monitor, delay and enforce safety guidelines

[23] proposes general policy desiderata for a world about to countenance AGI/ASI and recommend with specific reference to the problem of potential global coordination failure, that control of such technology be centralized, or a monitoring regime devoted to identifying harmful applications of AI and

intercepting them pre-deployment. As seen above, creating a global AGI project would address the global coordination issue head-on. Achieving a level of coordination superior to that which currently sustains the International Space Station would be a precondition for an international AGI development program.

What about a world which is not yet ready for such coordination or the advent of AGI; namely, our world? We propose that a global watchdog agency, be created for the express purpose of tracking progress of AGI programs, state-funded as well as corporate (the third variety, rogue AGI development, may be harder to monitor, but not impossible). This agency would have as its operating charter, the treaty on safe AGI development for the benefit of all humanity, as proposed above, with jurisdiction across all nations (a singleton, but within the specific context of AGI development, without authority over other aspects of governance or administration) and the lawful authority to both intercept unlawful attempts at AGI development and unilaterally terminate or freeze such programs.

The creation of such a body with such vast powers (albeit within a specific context), would also be constrained by the efficacy of global coordination and is less likely to happen relative to the space-station analogue proposed earlier. It is perhaps easier to bring people together to do something (build AGI as a global effort), rather than to create a group that tells everyone else not to do specific things (watchdog with the power to stop AGI development that runs afoul of the treaty). The history of the IAEA may be of interest if this path is chosen, to learn from its successes and preempt challenges likely to be faced by an international agency tasked with regulating technological development of a specific sort. Any such agency would benefit from drawing upon the cybersecurity infrastructures established by nation states, as well as the intelligence communities of the participating states. The proposed agency would, however, have a focus that is global, and not constrained by the national interests of participating nations, as this could easily lead to conflicts with undesirable ramifications.

Now, there are reasons to believe that AGI development may occur in the corporate sphere, as opposed to within a state-funded program. OpenAI is a non-profit company formed recently by a number of entrepreneurs concerned about safe AGI development. Several companies such as Google DeepMind, Vicarious, etc. are pushing the boundaries of what AI is capable of, with an increasing rate of progress. Many such companies could form a consortium driven by the need for safe AGI development and public good. Indeed, such an entity exists in the form of Partnership on AI [24] a timely non-profit organization bringing together diverse parties and actors with public safety and benefit as foci.

The problem with this scenario is that the raw capability for AGI development is not strictly limited to one corner of the globe, and in the absence of intergovernmental coordination and a global regulatory authority with real legal power to halt unsafe programs, it is entirely conceivable that AGI development could occur in multiple parts of the world and not all players may accept the rules likely to ensure the safest and most beneficial outcome. [25] proposes the creation of an AI Standards Developing Organization, whose role would be to provide strict guidelines for risk management and AI safety in an industrial context. This is indeed a good non-state analogue to the adoption of a treaty, and perhaps closer to the theater of AI development/deployment. Although it is to be expected that complexities associated with developing regulatory standards would necessitate the involvement of states or a coalition of states.

In addition to the solutions proffered above, powerful impetus could be provided to the creation of a Nanny AI [26] which would be tasked with the monitoring of AGI development worldwide with a clear mandate to delay any and all such programs until the coordination issues and safety considerations can be addressed rigorously. This, however, is not without risks of its own.

6. CONCLUSION

An artificial intelligence arms race most likely cannot be stopped, only managed. While there are many possible scenarios and outcomes, it is in the best interest of humanity that the dangerous ones be given due consideration before we develop AGI. We believe a systematic and tempered, public elucidation of the risks would help the cause of safe AGI development more than an approach characterized by hype and apocalyptic messaging. Technological progress tends to have a life of its own, and given the rate at which AI systems are achieving feats of intelligence and expertise, it is merely a matter of time, perhaps a few decades hence, perhaps more, before a truly general AI comes into existence. In this paper we have examined the prospect of such AGI development being prosecuted as an “arms race”, and have offered

a set of solutions, including the development of a comprehensive treaty on AGI development, international collaboration on a singular AGI program, a regulatory global watchdog designed to enforce the aforementioned treaty and potential recruitment of a Nanny-AI system [26], to delay AGI development until pragmatic considerations and risks can all be addressed with sufficient rigor. With this, we seek to add to the emerging discussion on AI safety within the technology and policy communities, and hope that the ideas presented herein are investigated thoroughly with concrete application in mind.

ACKNOWLEDGEMENT

We thank Beth M.Barnes for valuable comments on this paper.

REFERENCES

[1] A. Levinovitz. “The mystery of Go, the ancient game that computers still can’t win.” Wired Magazine. Retrieved August 27, 2017, from https://www.wired.com/2014/05/the-world-of-computer-go/, 2014.

[2] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Pannershelvam, M. Lanctot, and S. Dieleman, “Mastering the game of Go with deep neural networks and tree search,” Nature, vol. 529 no.7587, pp. 484-489, 2016.

[3] E. Gibney, "What Google’s winning Go algorithm will do next," Nature, vol. 531 no. 7594, pp. 284-285, 2016.

[4] E. Gibney, "Google secretly tested AI bot," Nature, vol. 541, pp. 142-142, 2017.

[5] N. Ernest, D. Carroll, C. Schumacher, M. Clark, K. Cohen, and G. Lee, "Genetic Fuzzy based Artificial Intelligence for Unmanned Combat Aerial Vehicle Control in Simulated Air Combat Missions," J Def Manag, vol. 6 no. 144, pp. 2167-0374, 2016.

[6] G. Booch, "I, for One, Welcome Our New Computer Overlords," IEEE Software, vol. 32 no. 6, pp. 8-10, 2015.

[7] K. Grace, J. Salvatier, A. Dafoe, B. Zhang, and O. Evans, "When Will AI Exceed Human Performance? Evidence from AI Experts," arXiv preprint arXiv:1705.08807, 2017.

[8] "Preparing for the future of artificial intelligence" The White House. Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf, 2016

[9] G. Allen, and T. Chan, "Artificial Intelligence and National Security," Retrieved from http://www.belfercenter.org/sites/default/files/files/publication/AI%20NatSec%20-%20final.pdf, 2017.

[10] S. Armstrong, N. Bostrom, and C. Shulman, "Racing to the precipice: a model of artificial intelligence development," Technical Report #2013-1, Future of Humanity Institute, Oxford University, 2013.

[11] C. Von Clausewitz, "On war: translated from the German by OJ Matthijs Jolles," The Modern Library, 1943.

[12] R. Yampolskiy, "Leakproofing the singularity: artificial intelligence confinement problem," Journal of Consciousness Studies, vol. 19 no. 1-2, pp. 194-214, 2012.

[13] R. Feynman, "Report of the presidential commission on the space shuttle Challenger accident. Appendix F," Retrieved from https://engineering.purdue.edu/~aae519/columbialoss/feynman-rogersrpt-app-f.pdf, 1986

[14] B. Watson, “The Drones of ISIS,” Defense One. Retrieved, August 26, 2017 from http://www.defenseone.com/technology/2017/01/drones-isis/134542/, 2017

[15] F. Pistono, and R. Yampolskiy, “Unethical research: how to create a malevolent artificial intelligence,” arXiv preprint arXiv:1605.02817, Retrieved August 27, 2017, from https://arxiv.org/abs/1605.02817, 2016

[16] S. Russell, D. Dewey, and M. Tegmark, "Research priorities for robust and beneficial artificial intelligence," Ai Magazine, vol. 36 no. 4, pp. 105-114, 2015.

[17] S. D. Baum, "On the promotion of safe and socially beneficial artificial intelligence," AI & Society, pp. 1-9, 2016.

[18] N. Bostrom, "Strategic implications of openness in AI development," Global Policy, vol. 8 no. 2, pp. 135-148, 2017.

[19] A. Critch, "Toward negotiable reinforcement learning: shifting priorities in Pareto optimal sequential decision-making," arXiv preprint arXiv:1701.01302, 2017.

[20] G. H. Kitmacher, W. H. Gerstenmeier, J. Bartoe, and N. Mustachio, "The international space station: A pathway to the future," Acta astronautica, vol. 57 no. 2, pp. 594-603, 2005.

[21] https://futureoflife.org/2017/08/20/leaders-top-robotics-ai-companies-call-ban-killer-robots/

[22] https://futureoflife.org/ai-principles/

[23] N. Bostrom, A. Dafoe, and C. Flynn, "Policy Desiderata in the Development of Machine Superintelligence," Working Paper, Future of Humanity Institute, Oxford University, 2016.

[24] https://www.partnershiponai.org/#

[25] S. Ozlati, and R. Yampolskiy, "The Formalization of AI Risk Management and Safety Standards," 31st AAAI Conference on Artificial Intelligence (AAAI-2017). 3rd International Workshop on AI, Ethics and Society. San Francisco, CA, USA. 2017.

[26] B. Goertzel, "Should Humanity Build a Global Nanny AI to Delay the Singularity Until It's Better Understood?" Journal of Consciousness Studies, vol. 19 no. 1-2, pp. 96-111, 2012.