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

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

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.