Article

Avoiding AGI races through self-regulation

G Gordon Worley III [1,*]

[1] Phenomenological AI Safety Research Institute; gordon@paisri.org
* Correspondence: gordon@paisri.org; Tel.: +1-507-697-5293

Abstract: The first group to build artificial general intelligence or AGI stands to gain a significant strategic and market advantage over competitors, so companies, universities, militaries, and other actors have strong incentives to race to build AGI first. An AGI race would be dangerous, though, because it would prioritize capabilities over safety and increase the risk of existential catastrophe. A self-regulatory organization (SRO) for AGI may be able to change incentives to favor safety over capabilities and encourage cooperation rather than racing.

Keywords: artificial general intelligence, AI policy, self-regulatory organization)

1. Introduction

The history of modern technology has often been a history of technological races. A race starts when a new technology becomes cost-effective, then companies, states, and other actors hurry to develop the technology in hopes of capturing market share before others can, gaining a strategic advantage over competitors, or otherwise benefiting from a first-mover position [1]. Some notable examples of recent technological races include races over rockets, personal computers, and DNA sequencing [2], [3], [4]. Although most of these races have been generally beneficial for society by quickly increasing productivity and expanding the economy, others, like races over weapons, generally make us less safe. In particular the race to build nuclear weapons dramatically increased humanity's capability to extinguish itself and exposed us to new existential risks that we previously did not face [5], [6]. This means that technological races can harm as much as they can help, and nowhere is that more true than in the burgeoning race to build AI [7].

In particular we may be near the start of a race to build artificial general intelligence or AGI thanks to recent advances in deep learning [8]. And unlike existing narrow AI that outperforms humans but only on very specific tasks, AGI will be as good or better than humans at all tasks such that an AGI could replace a human in any context [9]. The promise of replacing humans with AGI is extremely appealing to many organizations since AGI could be cheaper, more productive, and more loyal than humans, so the incentives to race to build the first AGI are strong [10]. Yet the very capabilities that make AGI so compelling also make them extremely dangerous, so we may find we would have been better off not building AGI at all if we cannot build them safely [11].

The risks of AGI have been widely discussed, but we may briefly summarize them by saying AGI will eventually become more capable than humans, AGI may not necessarily share human values, and so AGI may eventually act against humanity's wishes in ways that we will be powerless to prevent [12], [13]. This means AGI presents a new existential risk similar to but far more unwieldy than the one created by nuclear weapons, and unlike nuclear weapons that can be controlled with relatively prosaic methods, controlling AGI demands solving the much harder problem of value aligning an "alien" agent [14], [15], [16]. Thus it's especially dangerous if there is a race for AGI since it will create incentives to build capabilities out in advance of our ability to control them due to a likely tradeoff between capabilities and safety [17]. This all suggests that building safe AGI requires in part resolving the coordination problem of avoiding an AGI race. To that end we consider the creation of a self-regulatory organization for AGI to help coordinate AGI research efforts to ensure safety and avoid a race.

 doi:10.20944/preprints201810.0030.v1

2. An SRO for AGI

Self-regulatory organizations (SROs) are non-governmental organizations (NGOs) setup by companies and individuals in an industry to serve as voluntary regulatory bodies. Although they are sometimes granted statutory power by governments, usually they operate as free associations that coordinate to encourage participation by actors in their industries, often by shunning those who do not participate and conferring benefits to those that do [18]. They are especially common in industries where there is either a potentially adversarial relationship with society, like advertising and arms, or a safety concern, like medicine and engineering [19]. Briefly reviewing the form and function of some existing SROs:

- TrustArc (formerly TRUSTe) has long provided voluntary certification services to web companies to help them assure the public that they are following best practices that allow consumers to protect their privacy. TrustArc has been successful enough to, outside the EU, keep governments from much regulating online privacy issues [20].
- The US Green Building Council offers multiple levels of LEED certification to provide both targets and proof to the public that real estate developers are protecting environmental commons [21].
- The European Adversing Standards Alliance and the International Council for Ad Self-Regulation encourage advertisers to self-regulate and adopt voluntary standards that benefit the public to avoid the imposition of potentially less favorable and more fractured governmental ad regulation [22].
- The American Medical Association, the American Bar Association, the National Society of Professional Engineers, and the National Association of Realtors are SROs that function as de facto official regulators of their industries in the United States. They act to ensure doctors, lawyers, engineers, and realtors, respectively, follow practices that serve the public interest in the absence of more comprehensive government regulation [23].
- Although governments have progressively taken a stronger hand in financial regulation over the past 100 years, many segments of the financial industry rely in part on SROs to shape their actions and avoid unwanted legislative regulation [24].

Currently computer programmers, data scientists, and other IT professionals are largely unregulated except insofar as their work touches other regulated industries [25]. There are professional associations like the IEEE and ACM and best-practice frameworks like ITIL, but otherwise there are no SROs overseeing the work of companies and researchers pursuing either narrow AI or AGI, yet as outlined above narrow AI and especially AGI are areas where there are many incentives to build capabilities that may unwittingly violate societal preferences and damage the public commons. Consequently, there may be reason to form an AGI SRO. Some reasons in favor:

- An SRO could offer certification of safety and alignment efforts being taken by AGI researchers.
- An SRO may be well positioned to reduce the risk of an AGI race by coordinating efforts that would otherwise result in competition.
- An SRO could encourage AGI safety in industry and academia while being politically neutral (not tied to a single university, company, or nation).
- An SRO may allow AGI safety experts to manage the industry rather than letting it fall to other actors who may be less qualified or have different concerns that do not as strongly include prevention of existential risks.
- An SRO could act as a "clearinghouse" for AGI safety research funding.
- An SRO could give greater legitimacy to prioritizing AGI safety efforts among capabilities researchers.

Some reasons against:

- An SRO might form a de facto "guild" and keep out qualified researchers.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 October 2018 doi:10.20944/preprints201810.0030.v1

		Company A: First	Company A: Last
Company B	First	$(1,1)$	$(-1,3)$
	Last	$(3,-1)$	$(1,1)$

Table 1. Payout matrix for competing to build AGI

		Company A: Races	Company A: Cooperates
Company B	Races	$(1,1)$	$(-1,3)$
	Cooperates	$(3,-1)$	$(1,1)$

Table 2. Payout matrix for AGI race game

- An SRO could create the appearance that more is being done than really is and thus disincentivize safety research.
- An SRO could relatedly promote the wrong incentives and actually result in less safe AGI.
- An SRO might divert funding and effort from technical research in AGI safety.

On the whole this suggests an SRO for AGI would be net positive so long as it were well managed, focused on promoting safety, and responsive to developments in AGI safety research. In particular it may offer a way to avoid an AGI race by changing incentives to avoid the game theoretic equilibriums that cause races.

3. Using an SRO to Reduce AGI Race Risks

To see how an SRO could reduce the risk of an AGI race, let's first consider a game theoretic description of the AGI race.

Suppose that there are two entities trying to build AGI — company A and company B. It costs \$1 trillion to develop AGI, a cost both companies must pay, and the market for AGI is worth \$4 trillion. If one company beats the other to market it will capture the entire market thanks to its first-mover advantage, netting the company \$3 trillion in profits, and the company that is last to market earns no revenue and loses \$1 trillion. If the companies tie, though, they split the market and each earn \$1 trillion. This scenario yields the the payout matrix in Table 1.

The scenario is symmetric, the expected value of trying to win is $0.5(-1) + 0.5(3) = 1$, the expected value of tying is $0.5(1) + 0.5(1) = 1$, and the expected value of competing is $0.25(1) + 0.25(1) + 0.25(-1) + 0.25(3) = 1$, thus companies A and B should be indifferent between trying to win and tying. Given this it seems it should be easy to convince both companies that they should cooperate for a tie and coordinate their efforts so that they can focus on safety, but this immediately creates a game where each company must choose whether to honestly cooperate or pretend to cooperate and race in secret [26]. If both race or both cooperate their expected values remain 1, but if one races and the other cooperates then the racer stands to win at the expense of the cooperator. The payout matrix for this game is given in Table 2.

We see that the AI race game is symmetric, the expected value of racing is $0.5(1) + 0.5(3) = 2$, and the expected value of cooperating is $0.5(-1) + (0.5)1 = 0$, so it seems both companies should be inclined to race lest they lose by cooperating when the other company races, and an easy way to get ahead in the race is to ignore safety in favor of capabilities. What this game ignores, of course, is the large externalities imposed upon humanity by the development of unsafe AGI, especially the negative externalities of ignoring safety [27]. If we were to incorporate both the large, positive expected value of developing safe AGI that might be developed in the cooperate-only case and the large, negative expected value of developing unsafe AGI in the race cases, the payout matrix for the game would more resemble that given in Table 3, making clear that the incentives in the race game are poorly aligned with the interests of humanity.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 October 2018 doi:10.20944/preprints201810.0030.v1

		Company A	
		Races	Cooperates
Company B	Races	$-\infty$	$-\infty$
	Cooperates	$-\infty$	∞

Table 3. Payout matrix for AGI race game with externalities

In order to encourage both companies to cooperate and thus be more likely to build safe AGI we want to look for interventions that transform the payout matrix of the game so that it more resembles Table 3 with externalities included and less resembles Table 2 where they are ignored. Many options are possible and some are explored elsewhere in this special issue, but one way to change the equilibrium of the AGI race game would be with an SRO for AGI which could impose externalities on the companies by various methods including:

- inspections to demonstrate to each company that the other is cooperating
- contractual financial penalties that would offset any gains from defecting
- social sanctions via public outreach that would reduce gains from defecting
- sharing discoveries between companies
- required shutdown of any uncooperatively built AGI

In this example we need penalties worth in excess of $2 trillion imposed on companies that race to make them prefer to cooperate, which in the real world would likely require the combination of several strategies to make sure the bar is cleared even if one or several sources of penalties fail. Some of these strategies may also require enforcement by state actors, which further complicates the situation since militaries may also be participating in the race, and suggests an SRO may be insufficient to prevent an AGI race unless it is partnered with an intergovernmental organization, such as the United Nations [28]. That said a more traditional SRO could act faster with fewer political entanglements, so there seems to be space for both an SRO focused on industrial and academic AGI research and an intergovernmental organization working in collaboration with it to adjust the incentives of state actors.

The key takeaway is that even if an SRO is not the best way to modify the equilibrium of the AGI race, there is a need for some organization to impose externalities that reduce the chance of an AGI race by making it less appealing than when externalities can be ignored. SROs provide a clear template for this sort of organization, though addressing the AGI race specifically may require innovative policy solutions outside of those normally taken by SROs. An SRO for AGI thus stands likely to be a key component in avoiding an AGI race if it is willing to evolve in ways that help it address the issue.

4. Conclusion

An SRO for AGI is likely valuable, and may be particularly helpful in counteracting the incentives to race to develop AGI. Although there is currently no SRO for AGI, there are several organizations that are already positioned to take on an SRO role if they so chose, although some better positioned than others. They include:

- Partnership on AI
- Centre for the Study of Existential Risk
- World Economic Forum Council on AI and Robotics
- International Telecommunications Union
- Future of Life Institute
- Future of Humanity Institute
- Leverhulme Center for the Future of Intelligence
- Machine Intelligence Research Institute
- Center for Human-Compatible AI
- Center for Safety And Reliability of Autonomous Systems

If none of these groups wish to take on the task then creating an SRO for AGI is likely a neglected cause for those concerned about the existential risks posed by AGI. It is the recommendation of the present work that either an existing organization or a new one take up the task of serving as an SRO for AGI to reduce the risk of an AGI race and otherwise foster safety in AGI research.

Funding: This research received no external funding.

Acknowledgments: Thanks to Qiaochu Yuan and John Maxwell IV for comments on the ideas that eventually lead to this paper.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AGI	Artificial General Intelligence
SRO	Self-regulatory Organization
NGO	Non-governmental Organization

References

1. Gottinger, H.W. *Innovation, Technology and Hypercompetition: Review and Synthesis*; Routledge Studies in Global Competition, Taylor & Francis, 2013.
2. Collins, M.J. *Space Race: The U.S.-U.S.S.R. Competition to Reach the Moon*; Pomegranate catalog, Pomegranate Communications, 1999.
3. Cringely, R.X. *Accidental Empires*; HarperCollins, 1996.
4. Davies, K. *Cracking the Genome: Inside the Race to Unlock Human DNA*; Johns Hopkins University Press, 2002.
5. Powaski, R.E. *Return to Armageddon: The United States and the Nuclear Arms Race, 1981-1999*; Oxford University Press, 2000.
6. Bostrom, N. Existential Risk Prevention as a Global Priority. *Global Policy* **2013**, *4*. Accessed on Mon, September 24, 2018.
7. Armstrong, S.; Bostrom, N.; Shulman, C. Racing to the Precipice: a Model of Artificial Intelligence Development. Technical report, 2013.
8. Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; Chen, Y.; Lillicrap, T.; Hui, F.; Sifre, L.; van den Driessche, G.; Graepel, T.; Hassabis, D. Mastering the game of Go without human knowledge. *Nature* **2017**, *550*, 354–359. doi:10.1038/nature24270.
9. Shoham, Y.; Perrault, R.; Brynjolfsson, E.; Clark, J.; LeGassick, C. 2017 Annual Report. Technical report, 2017.
10. Hanson, R. Economic Growth Given Machine Intelligence. Technical report, 2001.
11. Bostrom, N. Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. *Journal of Evolution and Technology* **2002**, *9*.
12. Müller, V.C. Risks of general artificial intelligence. *Journal of Experimental & Theoretical Artificial Intelligence* **2014**, *26*, 297–301. doi:10.1080/0952813x.2014.895110.
13. Brundage, M.; Avin, S.; Clark, J.; Toner, H.; Eckersley, P.; Garfinkel, B.; Dafoe, A.; Scharre, P.; Zeitzoff, T.; Filar, B.; Anderson, H.S.; Roff, H.; Allen, G.C.; Steinhardt, J.; Flynn, C.; hÉigeartaigh, S.Ó.; Beard, S.; Belfield, H.; Farquhar, S.; Lyle, C.; Crootof, R.; Evans, O.; Page, M.; Bryson, J.; Yampolskiy, R.; Amodei, D. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. Technical report, 2018.
14. of the Deputy Assistant Secretary of Defense for Nuclear Matters, O. Nuclear Matters Handbook 2016. Technical report, 2016.
15. Turchin, A.; Denkenberger, D. Classification of global catastrophic risks connected with artificial intelligence. *AI & Society* **2018**. doi:10.1007/s00146-018-0845-5.

16. Yudkowsky, E. AI Alignment: Why It's Hard, and Where to Start. Symbolic Systems Distinguished Speaker Series, 2016.
17. Worley III, G.G. Robustness to fundamental uncertainty in AGI alignment. Technical report, 2018. Accessed on Tue, September 18, 2018.
18. DeMarzo, P.M.; Fishman, M.J.; Hagerty, K.M. Self-Regulation and Government Oversight. *The Review of Economic Studies* **2005**, *72*.
19. Shuchman, H.L. *Self-regulation in the professions*; Futures Group, 1981.
20. Benassi, P. TRUSTe: an online privacy seal program. *Communications of the ACM* **1999**, *42*, 56–59. doi:10.1145/293411.293461.
21. Cidell, J. A political ecology of the built environment: LEED certification for green buildings. *Local Environment* **2009**, *14*, 621–633. doi:10.1080/13549830903089275.
22. European Commission. Self-Regulation in the EU Advertising Sector: A Report of Some Discussion Among Interested Parties. Technical report, 2006.
23. Freidson, E. *Professionalism Reborn: Theory, Prophecy, and Policy*; University Of Chicago Press, 1994.
24. DeMarzo, P.M.; Fishman, M.J.; Hagerty, K.M. Contracting and Enforcement with a Self-Regulatory Organization. *SSRN Electronic Journal* **2002**. doi:10.2139/ssrn.297302.
25. National Society of Professional Engineers. The Cheapening of the Engineer Title. Accessed on Wed, September 19, 2018.
26. Schelling, T.C. *The Strategy of Conflict*; Harvard University Press, 1981.
27. Bostrom, N.; Dafoe, A.; Flynn, C. Public Policy and Superintelligent AI: A Vector Field Approach. In *Ethics of Artificial Intelligence*; Oxford University Press, 2019.
28. Perkovich, G.; Mathews, J.T.; Cirincione, J.; Gottemoeller, R.; Wolfsthal, J.B. Universal Compliance: A Strategy for Nuclear Security. Technical report, 2007.