FORESIGHT INSTITUTE

Artificial General Intelligence: Coordination & Great Powers

A white paper based on the
2018 Foresight Institute Strategy Meeting on AGI

Allison Duettmann, Foresight Institute
Olga Afanasjeva, GoodAI
Stuart Armstrong, Future of Humanity Institute
Ryan Braley, Lightbend
Jessica Cussins, Future of Life Institute
Jeffrey Ding, Future of Humanity Institute
Peter Eckersley, Partnership on AI
Melody Guan, Stanford University
Alyssa Vance, Apprente
Roman Yampolskiy, University of Louisville

Artificial General Intelligence: Coordination & Great Powers

A white paper based on the 2018 Foresight Institute Strategy Meeting on AGI

Report authors:
Allison Duettmann, Foresight Institute
Olga Afanasjeva, GoodAI
Stuart Armstrong, Future of Humanity Institute
Ryan Braley, Lightbend
Jessica Cussins, Future of Life Institute
Jeffrey Ding, Future of Humanity Institute
Peter Eckersley, Partnership on AI
Melody Guan, Stanford University
Alyssa Vance, Apprente
Roman Yampolskiy, University of Louisville

Foresight Institute is a non-profit organization to steer the beneficial development of technologies of fundamental importance for the future of life, focused on AI, cybersecurity, and molecular machine nanotechnology. Foresight selectively advances beneficial technologies via technical workshops, the Feynman Prizes and the Foresight Fellowships, and avoids the risks of technologies via strategy meetings and policy recommendations.

Participants

Afanasjeva, Olga	GoodAI
Armstrong, Stuart	Future of Humanity Institute
Baum, Seth	Global Catastrophic Risk Institute
Belfield, Haydn	Centre for the Study of Existential Risk
Bensinger, Rob	Machine Intelligence Research Institute
Bourgon, Malo	Machine Intelligence Research Institute
Bowerman, Niel	80,000 Hours
Braley, Ryan	Lightbend
Brown, Tom	Google Brain
Burja, Samo	Bismarck Analysis
Carey, Ryan	Ought, 2018 Foresight Institute Fellow
Cooper, Betsy	Center for Long-Term Cybersecurity
Cuperman, Miron	Base Zero
Cussins, Jessica	Future of Life Institute
Ding, Jeffrey	Future of Humanity Institute
Duettmann, Allison	Foresight Institute
Eckersley, Peter	Electronic Frontier Foundation
Fischer, Kevin	Crypto Lotus
Garfinkel, Benjamin	Future of Humanity Institute
Guan, Melody	Dept. of Computer Science, Stanford
Havrda, Marek	GoodAI
Irving, Geoffrey	OpenAI
Kotran, Alex	The Future Society
Krakovna, Victoria	DeepMind
Kramar, Janos	DeepMind
Lai, Tony	Legal.io
Leung, Jade	Future of Humanity Institute
Liston, Matthew	ConsenSys
Maas, Matthijs	Global Catastrophic Risk Institute
Mallah, Richard	Future of Life Institute
Mangan, Fiona	Justice and Security in Transitions
McReynolds, Joe	Jamestown Foundation
Michaud, Eric	Rift Recon
Miller, Mark	Foresight Institute, Agoric
Mosleh, Ali	John Garrick Institute for the Risk Sciences
Nitzberg, Mark	Center for Human Compatible AI
O'Neill, Jim	Mithril Capital
Olsson, Catherine	Google Brain
Page, Michael	OpenAI
Peterson, Christine	Foresight Institute
Scheyer, Peter	2018 Foresight Institute Fellow
Shulman, Carl	Future of Humanity Institute
Singh, Tanya	Future of Humanity Institute
Snow, Rion	Unaffiliated
Tallinn, Jaan	Future of Life Institute
Vance, Alyssa	Apprente
Webb, Michael	Dept. of Economics, Stanford Univ.
Wu, Dekai	Hong Kong University of Science & Technology
Xiao, Qiang	School of Information, UC Berkeley
Xu, Mimee	UnifyID
Yampolskiy, Roman	University of Louisville, 2018 Foresight Fellow

Table of Contents

Executive Summary

Allison Duettmann
Foresight Institute

The annual Foresight Institute AGI strategy meeting gathers representatives of AI safety organizations and academic institutions with experts in fields relevant to AGI strategy, including security, government policy, and international political economy. The 2017 Foresight Institute AGI strategy meeting on AGI Timeframes & Policy focused on AI timelines, with special consideration given to policy, cybersecurity, and coordination. A 72% majority of workshop survey respondents voted for the 2018 AGI strategy meeting to focus on avenues for coordination on the path toward AGI, especially in relation to the world's greatest geopolitical powers.

This emphasis parallels a recent upturn in AI safety organizations' focus on coordination, governance and policy issues, suggesting that a meeting to summarize and share progress across organizations was timely. Some recently launched projects on public policy considerations around AI that our meeting participants are involved with include:

- 80000 Hours hired a specialist coach on AI policy, Niel Bowerman, who was present at the meeting.
- A joint CSER/CFI team collaborated with the UN's AI for Good Conference and the launch of the Ada Lovelace Institute. Haydn Belfied from the Centre for the Study of Existential Risk was present.
- GoodAI launched the Solving The AI Race Challenge. Olga Afanasjeva, the Director of the Challenge (remote), and Roman Yampolskiy, a Judge in the Challenge, were both present.
- DeepMind launched its Ethics & Society research unit, focusing on issues relating to privacy, transparency, economic impact, governance and accountability, and helped to start the AI Now Institute at New York University. Viktoria Krakovna of DeepMind participated remotely during parts of the meeting.
- Future of Humanity Institute launched its Governance of AI Program. Jade Leung, Ben Garfinkel, Tanya Singh, and Jeffrey Ding (remote) were present during the meeting.
- Future of Life Institute issued its 2018 grant recommendations, including recommending grants to workshop attendees Michael Webb on the topic of Transition to AI economy. Future of Life Institute also released a LAWS open letter with 2400 signatories. Richard Mallah and Jessica Cussins from Future of Life

Institute were present.

- OpenAI launched its Charter. Geoffrey Irving and Michael Page from OpenAI were present.
- Google announced its AI Principles. Tom Brown from Google Brain was present.
- The US Office of Science & Technology Policy issued the Preparing for the Future of AI Report, and launched a Select Committee on AI. Malo Bourgon and Rob Bensinger of the Machine Intelligence Research Institute, which submitted a response to the Request for Information for the report were present.
- The UK Parliament established a Select Committee on AI. Haydn Belfield of the Centre for the Study of Existential Risk and Peter Eckersley of the Electronic Frontier Foundation, whose organizations both submitted responses to the Request for Comments issued by the Committee, were present.

Making progress on AI safety requires making progress in several domains, including ethics, technical alignment, cybersecurity, and human coordination, all of which contain a number of hard problems. Ensuring coordination among actors that facilitates cooperation on solving those problems, while avoiding race-dynamics that may lead to cutting corners on safety issues, is a primary concern on the path to AI safety. While coordination is itself a very hard problem, making any threshold progress on coordination upfront would be beneficial for addressing ethics, technical alignment, and cybersecurity concerns by allowing more time to solve those issues. Since coordination for AGI safety can involve existing actors, and literature and historical precedents about dealing with similar coordination challenges are available to inform our approach, coordination is a goal that we can and should effectively work toward today. Current geopolitical developments, including dueling tariff proposals between China and the US, signs of potential resurgent nuclear proliferation and AI military arms race dynamics, only increase the urgency of working toward coordination.

Identifying potential avenues for AGI coordination among important global actors can create collateral advantages for coordination on other impending risks as well. While the meeting's focus remained on AGI coordination, most other anthropogenic risks, such as those arising from potential biotechnology weapons, require coordination as well. Thus, while most claims in this paper are AGI-specific, other more generic recommendations for coordination may provide useful starting points for creating an overall policy framework that promotes robustness, resiliency, or even antifragility.

By gathering multiple AI safety organizations to discuss AGI coordination, the meeting entitled A*GI: Coordination & Great Powers* was itself a useful exercise in coordination. Participating organizations shared progress on their work and explored potential avenues for new or further collaboration. Many insights arose from the multidisciplinary exchange among AI safety groups, policy sector, and security sector, e.g. on the socio-political constraints for technical safety research, and the alarmingly deficient state of security today. Ultimately, these meetings allow organizations to create an understanding of trust and converge on common norms, which may aid in avoiding negative Unilterateralist's Curse-style scenarios within the AI safety community (Bostrom, 2013).

This report approaches AGI in three sections: (1) frameworks for coordination, (2) coordination scenarios, especially among governmental actors, among military actors, and among private actors, and (3) technological factors that may influence coordination, including cybersecurity and blockchain. The chart below shows a brief summary of the preliminary recommendations and requests for further research contained in each chapter of the report.

For comments or questions about the workshop series or the report, please contact lead author, Allison Duettmann, at a@foresight.org

Executive Summary

Chapter	Recommendation	Further research
Frameworks for coordination	Encourage early cooperation on concrete issues with lower stakes to create precedent and infrastructure for later cooperation	Strategies to incentivize relevant actors effectively given that their motives for developing AGI may be different
	Create shared frameworks and curricula, supported by benchmarks commonly accepted among actors	
Coordination scenarios	Capacity-building of long-term community of experts and decision-makers based on trust, who can coordinate quickly if necessary	Investigate imperfect precedent case of nuclear weapons coordination, especially the degree to which individuals are willing to work on dangerous activities, the possibility of flagging dangerous trends, and the ability to intervene in effective ways, especially during crisis situations
AI Strategy of Governmental Actors	Increase collaboration among government, industry, and academia, e.g. governments investing in AI by focusing on Fairness, Accountability, and Transparency (FAT)	Research incentive structures that incentivize the incentivizer
	Prevent malicious use by non-state actors, e.g., via strategies similar to the current effort to prevent autonomous killer robots	Bridge national information gaps among technology creators and regulators, e.g., between the Bay Area and Washington, D.C.
	Changes to the immigration law focused on talent attraction	
Military Strategy and AI Coordination	Detailed discussions and cultural exchange among relevant actors on a personal level	The possibility to reframe a race in capabilities to a race in security, predictability and control
	Create cultural memes that support successful coordination	
AI Strategy of Major Private Actors	Signal cooperation via public statements, collaborative research, and leading by example	Rethink openness considerations in research, e.g., via novel information sharing regimes
Incentivizing Safety Research	Concretizing the safety research agendas to make safety research more tractable	Find partial solutions to existing safety problems to show that progress is possible
	Highlight security aspects of AI safety to increase interest in the issue at conferences and journals	
Cybersecurity	Encourage use of seL4 microkernel as operating system	Study potential future effects of quantum computing on cybersecurity
	Advocate for responsible disclosure of vulnerabilities by governments	
Blockchain	Outreach to goal-aligned individuals in the cryptocurrency community	Gain deeper understanding of the blockchain space in the Western and Asian context to investigate potential effects on AI safety. A few key factors that may serve as starting points are listed in the blockchain section
	Regulatory advocacy options to ensure potential future regulation is informed and sensible	

1. Frameworks for Coordination: Theory and History

Theoretical frameworks for coordination

Framing and limitations

This report is shaped inevitably by the particular perspectives of its authors, and that can introduce significant limitations on objective orientation given the generally Western backgrounds of the authors. The report investigates different coordination scenarios among a variety of actors on international, national, and individual levels. In all of those scenarios, framing effects are present: different governmental styles, cultural contexts, incentives, and worldviews can lead different actors to frame relevant factors for coordination in very different ways. A few potential framing effects observed at the meeting include:

- **Vocabulary**: Investigating coordination efforts to ensure that AI remains "good" requires an initial common understanding of what "good AI" means, which may vary among actors. For instance, defining "good" as "human-rights compliant" may reflect a Western framing and will likely differ from a Chinese definition, given that China views human rights differently from the West. Thus, finding definitions of the social good that are not framed in terms of individual rights may be useful if China is to be involved in coordination efforts.

- **Complexity**: Simplistic, stereotypical representations of other actors make an accurate depiction of the complexities of other actors difficult. For instance, the current US narrative often focuses on Beijing and depicts China as one closed entity pursuing a monolithic AI drive. However, as pointed out in China's AI Dream, the landscape of involved agendas, agencies, and actors is wide and complex. Understanding the details and nuances of another actor's situation is essential when developing realistic incentives for coordination (Ding, 2018).

This reports describes some potential framing effects pertaining to individual coordination scenarios upfront in each section, but it's likely that some additional effects have been missed, so further research is required. The Leverhulme Centre for the Future of Intelligence recently initiated a project on Global AI Narratives to investigate differences in AI narratives. A few avenues for addressing framing effects include: detailed, personal discussions among relevant international actors to establish the right context and avoid miscommunication, and investigation of specific signalling techniques available to actors that recognize the framing differences of other actors. Finding strategies that enable all actors to coordinate based on one common base of reality is important for successful coordination efforts.

1. Frameworks for Coordination: Theory and History

An overview of coordination

A myriad of factors are relevant for achieving successful AGI coordination. To better understand available pathways for coordination, several strategies have been employed by AGI safety organizations to obtain an overview of the space of coordination:

- **Clarifying definitions**: Different researchers and organizations use a variety of different terms to discuss concepts relating to Artificial Intelligence and Artificial General Intelligence. This linguistic heterogeneity may result in unnecessary disagreements or hide relevant disagreements among researchers and cause misunderstandings within the community that affect the media narrative, the public, and the political apparatus. For an overview of different AI definitions, including AI, AGI, and Superintelligence, see AGI Safety: Overview & Definitions (Duettmann, 2018). A widely accepted definition of AI-related concepts or clarifications in which definitions are used in a given context could enhance useful policy development and coordination by avoiding unnecessary confusion and miscommunication.

- **Measuring progress**: Specific coordination strategies will, at least partly, depend on AI capabilities at the time of the coordination effort. This makes measuring progress a useful goal for understanding the required coordination strategies. Promising efforts at measuring AI development include the AIindex, and the EFF's AI Progress Measurement. To summarize the EFF's AI Progress Measurement to date: recent progress in AI has been rapid. Certain advances have gotten a lot of attention. These include: reinforcement learning agents that play Go or Dota 2 better than humans can, high-quality image recognition systems, and "fake" speech and video synthesis that is increasingly difficult to distinguish from the real thing. Other profound forms of progress have been less discussed. Reading comprehension models can now read about as well as second-grade children. Techniques for automatically designing neural network architectures can solve supervised classification tasks as well as networks crafted by the best human experts, and those methods quickly advanced from requiring enormous computational resources to easy processing on a single GPU. Game-playing systems are demonstrating spontaneous acquisition of simple language from feedback. AI systems are outperforming prior algorithms for important technical tasks like compression and cache optimization. Deep networks are a tool for searching through arbitrary function spaces, and taken as a whole, the field's rapid and ongoing progress indicates that variants and combinations of them may well be sufficient for performing any sort of task that requires intelligence.

- **Roadmaps**: One strategy for gradually paving ways toward cooperation is to better understand the nature of potential AI races and explore solutions through roadmapping and prizes. Roadmapping can be instrumental in understanding the dynamics of races, by mapping involved actors and their interactions, and opening up novel viewpoints. Roadmaps can capture different scales of interactions (in the global arena or inside a group) and variations of races, such as races in narrow AI applications in various sectors, AI arms races, or Artificial General Intelligence races. For instance, GoodAI recently published this roadmap on AGI Races.

- **Prizes**: In addition to the roadmapping effort, Solving the AI Race challenge (part of the General AI Challenge series organized by GoodAI) incentivized people around the world with different backgrounds to address AI Race issues. While this effort doesn't provide a complete solution to a wide range of AI race issues, it is a valuable contribution to understanding sub-issues, such as the AI weapons race.

- **Analyzing the actors**: There are multiple options for categorizing the relevant actors in the AI-space, e.g., divisions according to nation states, type of organization, or ideological closeness. Since each division will lead to different models of coordination, it is instructive to consider the different types of actors and

potential race scenarios in more detail. In 2017, a survey was conducted of Artificial General Intelligence (AGI) research and development (R&D) projects (Baum 2017). The survey attempted to identify every project seeking to build AGI. While OpenAI and DeepMind are the two major projects with a stated goal of developing AGI, 45 total projects were identified. Just over half of these are based in the United States, and almost all are based either in the US or in country that is a US ally. A few projects are based in China or Russia, and each of these projects has strong ties to either the international academic community or Western entities. Most of the projects are found either in academia or at for-profit companies. Relatively few have military connections, and those that do are US academic projects with funding from DARPA or other military funders. This reality is consistent with military funding of academic projects across the AI field and does not appear to indicate any major strategic initiative by the US military. Finally, many of the projects were interconnected, either by deliberate collaboration or as a byproduct of sharing some of the same team members. These findings suggest a relatively cooperative AGI R&D landscape (Baum, 2017).

Recommendations

To ensure coordination on AGI, actors could encourage early cooperation on concrete issues with lower stakes to create precedents and infrastructure for later cooperation. One general strategy for building trust and cooperation is creating shared frameworks, curricula and benchmarks. As illustrated in the section on Incentivizing Safety later in this report, the creation of curricula encourages safety research by making research topics and results more tractable. Similarly, the creation of frameworks that are shared and agreed upon by different actors and supported by collective benchmarks for progress could produce a common foundation to support negotiation and cooperation among different key actors.

Further research

Given the diversity of potentially relevant actors for AGI coordination, further research is necessary to investigate how to engage relevant actors most effectively. Different motivations for creating AGI, such as commercial long-term profit, intellectual progress, and creating good in the world each require different incentives for coordinating action agendas. A comprehensive mapping of possible motivations of current actors would be useful in this regard. See below for a sample mapping of actors and their incentives, which could be 'fleshed out' to include specific actors and their particular incentives.

AI race dynamics

To investigate the likelihood that AI races will be pursued, it is instructive to compare prospective AI race dynamics with previous large scale private sector competitions (Baum, 2018) or previous governmental race dynamics, e.g. nuclear and space races. Ideologically, reframing the AI quest as a cooperative endeavor, rather than a race, could have a positive psychological and motivational effect on actors (Cave, 2018) and the potential for coordination among them. Historical examples of increased practical cooperation achieved through a focus on the potential positive effects of cooperation include the creation of CERN, ITER and the National Academy of Sciences. However, while avoiding adversarial language and focusing on the common good—a CERN for AI—could be helpful to achieving cooperation, it may not suffice to counteract the strong incentives to compete. Thus, comparing possible AGI coordination scenarios to historical cases of race dynamics is instructive for avoiding or managing potential AGI races.

1. Frameworks for Coordination: Theory and History

Type	Name
Actors	Companies
	Public-private partnerships (PPPs)
	Consortia
	AI Community
	Rogue Actors
	National Governments
	International Governments (EU, UN, etc.)
	Military and Security Departments
	Private Investors
	Individual Talent (finite pool of researchers)
	Media
	Non-governmental Organizations
	The Public
	Politicians
	Opinion Leaders
	Religious Groups
	AI team members
	Team leaders
Actor motivators	Knowledge (understanding the universe)
	Altruism
	Welfare
	Security
	Hunger for power
	Economic dominance
	Revenge
	Ego
	Political hegemony
	Team rivalry
	Mistrust inside teams
Actions	Espionage
	Hype
	Regulation and Control
	Collective (distributed) actions/processes vs. one party actions
	International tensions and conflicts
	Public perception, global
	Internal state affairs
	Technical barriers
	Global disasters and slow down
	Safety barriers
	Progress on AGI
Interventions on processes	Prioritize AGI
	Turning rogue
	Going undercover
	Monitoring progress of groups and talent
	Pull of funding
	Ethical restrictions
	Regulation and enforcement
	Restricted collaboration
	Theft
	Nationalization
	Change in public perception
	Change in AI community perception
	Manipulation of opinion
AI Race Drivers and Disruptors	Single partnership
	Multiple partnerships
	Nationalizing companies
	Information sharing
	Surveillance

A sample modeling of actor incentives and decision-dilemmas by GoodAI. For a detailed explanation of the incentives above, see Avoiding the Precipice.

Comparing space, nuclear, and AGI races

An AGI race can be described as a phenomenon where stakeholders compete or are driven to be the first to develop and deploy an AGI, which in turn would give them a strategic advantage. Given the danger that an AGI might be deployed by bad actors, it seems vital to find robust mechanisms to avoid races before they start, or at least to influence the creation and handling of AGI according to best safety practices. Achieving perfect coordination among all actors (individuals, corporations, nation states, etc.) seems unlikely, unless a robust set of incentives to cooperate is found. Traditional economic incentives might be insufficient compared to the unprecedented advantages AGI might bring, and a motivation to benefit humankind through technology might not appeal to all actors. This section compares potential AGI races with historical precedents of the space race and the nuclear arms race (for current information on catastrophic risk arising from nuclear weapons, see Baum and Barrett (2018) and Baum et al. (2018)). While there are some similarities, several features of the nature of the AGI coordination scenario suggest that an AGI arms race may be harder to manage than were the nuclear and space races.

	AGI vs. nuclear race	AGI vs. space race
Incentives	One difference between potential AGI race dynamics and past nuclear race dynamics is that the expected result of a nuclear attack would be a strong deterrent to allowing coordination to falter: the potential outcome of extinguishing millions of lives with a nuclear attack is rarely an intrinsically attractive prospect to the actor launching the attack. Given AI's promise of profitability for the private sector, the incentives for AGI proliferation are probably stronger than those for nuclear proliferation.	Similar to a potential AGI race, the historical race to space was in part driven by intrinsic incentives to develop space technologies, which existed at least somewhat independently of race dynamics. However, while the potential outcome of the space race for the losing party was negative, owing to a loss of national pride and military advantages, that was not as potentially catastrophic an outcome as could result in the case of falling behind on AGI.

	AGI vs. nuclear race	AGI vs. space race
Capabilities	Another difference between nuclear and AGI proliferation is the democratization of capability. Nuclear proliferation is expensive, limiting the number of realistic actors to national governments. While the cost of building AGI remains unknown and is potentially very high, AI development costs are falling and the field is becoming increasingly democratized. Many AI models are being developed and shared openly for the most part, allowing their use by a great number of actors, who thereby obtain access to distributed computing capacity eliminating the need to rely on expensive server farms.	Given the immense capital requirements and lack of industry engagement, private corporations were not plausible players in the early space race. While some corporations are gradually opening up access to space, the resource requirements for building rockets function as a gatekeeper that restricts the number of relevant actors in the space industry. Additionally, the International Space Station is serving as great attractor for cooperation, because joint projects can often outspend non-participants. In contrast, it is unlikely but not impossible that in the case of AGI, the resources necessary to create AGI could be acquired by even small-scale actors, decreasing the incentive to join coalitions.
Monitoring	A difference among nuclear, space, and AGI proliferation scenarios is the ease of monitoring. In the nuclear case, there are distinct resource needs for weapons enrichment that are distinguishable from civilian applications. In addition, nuclear sites are large-scale enough to allow for satellite-monitoring. In the case of AGI monitoring, dangerous AGI applications cannot as easily be distinguished from beneficial, civilian uses, and given the uncertainties around hardware requirements for AGI, it is questionable whether it is even possible to reliably monitor all AGI-developing actors in any meaningful way.	As with AGI vs. nuclear race: rocket-sites, similar to nuclear sites, are large-scale enough to allow for satellite-monitoring, while the potential to reliably monitor AGI development remains questionable due to the required resources.

Interplay of risk scenarios

Rather than limiting one's analysis to investigating parallels among risks that arise from nuclear weapons and AGI, it is further instructive to examine the effects of those risks on each other. In the short term it is possible that AGI development may exacerbate the risk of conflict and war. In the current actor landscape, the closer one actor is to approaching the development of AGI, the higher other actors' incentives become to prevent that actor from deploying AGI—potentially at all costs. Even if an actor is only suspected to be nearing the deployment of a system that is sufficiently strong to execute a strategic takeover and confer immense power onto the actor controlling the AGI, other adversarial actors have a strong incentive to avoid this takeover from happening. Even if the AGI-developing actor assured other actors that its AI was safe and its goals were constructed to benefit all of humanity, this claim is hard to prove, not only technically -- but also in the face of existing differences in underlying core values. Thus, while in itself posing an existential risk, AGI could potentially exacerbate the risk of war, nuclear extinction, or other catastrophic and existential risks even further.

Recommendations

To grasp the magnitude of the challenge posed by coordination on AGI development and safety in the context of likely scientific competition among global leaders, comparing the situation with past experiences around the space race and nuclear arms race yields helpful insights. In light of the heightened and compounded global risk levels attendant to any race to AGI, depending on various scenarios and actors involved creating a tight community of experts and decision-makers based on mutual trust and collaboration would be beneficial.

For example, in Beyond Mad?: The Race for Artificial General Intelligence, Ramamoorthy and Yampolskiy consider the possibility of an AGI arms race and propose solutions aimed at managing the development of such an intelligence without increasing the risks to the global stability and safety (Ramamoorthy, Yampolskiy). The paper reviews actors likely to be involved in the AGI race (state, corporate, and rogue actors). Among the proposed solutions are global collaboration on AGI development and safety among leading industrial nations under the umbrella of the UN, via a proposed Benevolent AGI Treaty. Enforcement may require support from a new agency, like a Global Task Force, to monitor and enforce safety guidelines on AGI research around the world. The global task force called for in Beyond Mad has similarities to a strategy discussed at the Foresight Institute 2017 AGI Strategy Meeting:

> "The idea of creating a global leadership council on AI safety, e.g., in the shape of a new governance board with representation of all affected parties has been proposed by Sam Altman and others (Bostrom, Dafoe, Flynn, 2017). Such a council could take current examples of tools for international cooperation as role model. The UN allows for many small conversations to be had at a high-level, which is important, given that the AI safety problem consists of evergrowing sub-domains, e.g., coordination, alignment, ethical considerations, and cybersecurity. A further example, smaller in scale but closer to AI safety, is the IEEE Global Initiative for Autonomous and Intelligent Systems. A problem that faces the UN, and could present itself in the AI safety council as well, is that without an efficient decision-making mechanism and sufficient incentives to abide to the decisions made at such councils, the resolutions lack executed force. [...]. To overcome this vagueness one could concretize permitted AI levels to agree on regulation: Although it is hard to distinguish which architectural features are more risky than others for AI safety, one could start to classify AI levels into categories, potentially barring recursive self-improvers, and advanced hardware. However, while hardware is relatively easy to monitor, many of the software constraints are hard to monitor and enforce (Duettmann, 2017)."

Further research

The imperfect historical precedent of global coordination around nuclear weapons requires further analysis. Special focus should be placed on the anticipation of a race, the degree to which individuals are willing to work on dangerous activities, the possibility of flagging dangerous directions and trends, and the ability to intervene in effective ways, especially during crisis situations. Those would be valuable factors in creating plans today that, if put in place 5-10 years in advance of a potential power struggle around AGI, could shape the course of action toward coordination and bend the curve toward better global outcomes.

2a. AGI Coordination Scenarios: Governmental Actors

AI Strategy of China and the US

Framing effects

When considering governmental actors, framing effects may exist not only on an international level among different actors, but also on a national level among different political parties or administrations. In US politics, ideas are usually framed in terms of market-driven vocabulary, even while discussing AI strategies and regulations in China, which are potentially better understood in the context of policy-driven governance. Conversely, the Chinese discourse about US AI strategies may underestimate market dynamics at play. Framing differences related to AI can be also observed over time within the US executive branch, e.g., between the Trump administration and the Obama administration. The Obama administration's AI approach included calls for "aggressive" public policy in relation to AI. In contrast to that approach, when announcing the formation of a committee on AI, the Trump administration claimed that it is "*not in the business of conquering imaginary beasts. We will not try to "solve" problems that don't exist. To the greatest degree possible, we will allow scientists and technologists to freely develop their next great inventions right here in the United States. Command-control policies will never be able to keep up. Nor will we limit ourselves with international commitments rooted in fear of worst-case scenarios.*" While this statement allows for several interpretations, it promises a generally hands-off approach to regulation of AI and alludes to the possibilities of global arms races. Since the framing of governmental approaches can lead to large-scale misunderstanding, it is an important factor to be addressed in the quest for coordination.

National AI strategies

In the US context, a 2017 JASON DoD study found that it was too early to determine relevant government action in regard to AGI. Recently a variety of national AI efforts have been launched and although most of these initiatives do not specifically address AGI as a goal, they are promising rapid progress on AI. In October and December of 2016, the Obama administration published three government reports on the management and development of AI: "Preparing for the Future of Artificial Intelligence," "The National Artificial Intelligence Research and Development Strategic Plan," and "Artificial Intelligence, Automation, and the Economy." These reports did not add up to a nation-

al AI strategy for the United States, but they did help jump-start the conversation about government-led guidance of AI. Numerous countries around the world have subsequently developed national strategies for managing and directing the development of AI. These include Canada's Pan-Canadian Artificial Intelligence (AI) Strategy; China's " New Generation Artificial Intelligence Development Plan"; France's "AI for Humanity"; India's "National Strategy for Artificial Intelligence #AIforAll"; Japan's "Artificial Intelligence Technology Strategy"; Singapore's AI Singapore program; South Korea's "Managing the Fourth Industrial Revolution" report; the UAE's Strategy for Artificial Intelligence ; and the UK's Sector Deal for AI. At least 15 additional countries, including the US, Mexico, Germany, and Australia, are actively exploring AI policies, initiatives, and strategies of various kinds.

Many of these governmental initiatives prioritize national leadership in at least certain aspects of AI development. There are however, numerous efforts towards intergovernmental cooperation as well. For example, The European Commission published a communication that outlines the European approach to AI, which includes a declaration of cooperation among European countries. Additionally, in June 2018 leaders of the G7 Summit committed to the "Charlevoix Common Vision for the Future of Artificial Intelligence." Ahead of the G7 meeting, the Canadian Prime Minister and French President also announced the creation of an international study group for AI. These efforts may still primarily reinforce national interests and remain housed within existing political entities, but they highlight awareness of the global scope of these technologies and their implications. The willingness to collaborate across borders is likely to be increasingly important in mitigating AI race dynamics globally.

National Approaches to Ethics

While concerns related to the ethics of AGI have not reached national levels yet, several governments have presented national AI policies and strategies that highlight their awareness of ethical concerns arising in the use of AI (Guan, M.Y. 2018):

- The UK's strategy specifically "consider[s] the economic, ethical and social implications of advances in artificial intelligence" and recommends preparing for disruptions to the labor market, open data and data protection legislation, data portability, and data trusts. It notes that "large companies which have control over vast quantities of data must be prevented from becoming overly powerful."
- France's strategy similarly includes a focus on developing an ethical framework for "inclusive and diverse AI" and avoiding the "opaque privatization of AI or its potentially despotic usage."
- India's strategy highlights the importance of AI ethics, privacy, security and transparency as well as the current lack of regulations around privacy and security.
- Canada has a National Cyber Security Strategy for protecting Canadians' digital privacy, security and economy and a commitment to collaborate with France on ethical AI.
- China has a National Standard on Personal Data Collection which addresses issues similar to those in the European Union's General Data Protection Regulation (GDPR). The nation's "New Generation Artificial Intelligence Development Plan" underlines the need to "strengthen research and establish laws, regulations and ethical frameworks on legal, ethical, and social issues related to AI and protection of privacy and property."
- China, Japan, and Korea have all recently revised their legislation on personal information protection, and France and Japan have formulated personal information protection rules for new industries such as cloudcomputing.
- The European Union Legal Affairs Committee recommends "privacy by design and privacy by default, informed consent, and encryption, as well as use of personal data need to be clarified."

2a. AGI Coordination Scenarios: Governmental Actors

Although this list may seem encouraging, it is important to remember that a government's promises and its real-world behavior can substantially diverge. For example, China is listed above in connection with positive statements on privacy and personal information protection, but in China and to a lesser extent the US, the national government is widely assumed to collect substantial information on citizens, regardless of public pronouncements to the contrary.

Zooming in on China and the US

The The Future of Life Institute recently published a digest of the global landscape of national and international AI strategies, outlining governmental AI strategies for more than 22 countries with the goal of promoting cooperation and coordination among countries and developing best practices. In contrast to this broad approach, the current report focuses on Chinese and US actors as hypothetical or test cases for potential coordination scenarios among great powers. While Western observers tend to focus on drawbacks in the Chinese approach, it does have some advantages for AI governance that should be taken into account by US-originating coordination efforts. Areas of contrast include:

	China	US
Recent national developments	China has become an important actor in governing risks associated with AGI, as illustrated in depth in Deciphering China's AI Dream (Ding, 2018). In July 2017, China's State Council, the country's cabinet body, issued an AI Development Plan that set a benchmark of USD 1.5 trillion for the scale of China's AI industry in 2030—a figure that would put China into a world-leading position. This goal, while ambitious, is not outside the realm of possibility. Across many drivers of AI development —including hardware, data, research talent and AI firms—China is making enormous progress. Alongside the growth of China's commercial AI ecosystem, some Chinese scholars and policymakers have paid increasing attention to issues of AI governance, with an emphasis on near-term issues such as producing reliable and controllable AI technology that meets certain technical standards. A landscape map of the views of Chinese AI researchers on the risks of AGI shows the mainstream view that AGI is too distant on the time horizon to merit substantive consideration of its unique risks, though some well-known Chinese AI researchers have advocated against the development of AGI because of its risks.	Most US discussions of AI politics and coordination assume, by default, that the American government will plan and implement some particular AI policy in the near future. The focus is generally on what this policy will look like, and how it might be modified through lobbying. However, since roughly the 1970s, there has been a decline in the US federal government's ability to implement any kind of policy (regardless of content or subject matter). This can be illustrated by the small number of bills passed by Congress (Cillizza, 2014), or the percentage of spending bills passed by the nominal Oct. 1st deadline, or the number of motions for cloture. A popular recent excuse for inaction has been that the American system cannot make significant changes without unified party control, but this ignores the facts that, for example, all Reagan-era bills passed through a Democrat-controlled House of Representatives. Hence, the best bet for the near future is that the nominal major decision-makers (Congress, White House, courts) will largely ignore the issue, and that what discussion does take place will have only marginal practical effects. This does not imply that the US will cease to be a factor—it is, after all, currently the world center of AI—but that state actions will be mostly determined by a combination of past policies (e.g., existing US military doctrine) and the personal priorities of individual leaders in the executive branch.

2a. AGI Coordination Scenarios: Governmental Actors

	China	US
Technology understanding	Historically, China has been relatively technocratically led, both in terms of political leaders and civil servants, with top leadership that typically has been educated at engineering schools. The military and civil ecosystem are fused around national priorities under an 'informatization' approach, including close integration of academia, industry, military, and a number of public/private/academic partnerships.	While the US may lack integration and technical understanding at the top, program managers of certain agencies generally have a good technical understanding of their focus area. However, those managers do not typically draft policies, and often return to industry eventually. While part of DARPA's mission is to investigate AI goals that are more long-term than those in the commercial sphere, in practice the research focus of projects is often more short-term due to organizational incentives, with some managers willing to forego paradigm-shifting basic research for in favor of near-term incremental improvements.
Policy testing	In some ways, the Chinese government is better equipped to address certain challenges in innovation because it can make quick executive decisions, e.g., trialing fully autonomous driving zones within a couple of years, likely half a decade ahead of the US.	By the time the US will have to engage with difficult policy and societal conversations about the liability, ethics and safety of autonomous driving zones, China will have been navigating this challenge for around 5 years. This lag time can present an opportunity for the US to learn from the Chinese case. Structuring coordination conversations with China around these finite, manageable challenges that both countries are aligned on can be useful to establishing provisional bridges for future coordination.
Immigration	China historically has not issued passports to non-Chinese residents. However, China is recently exploring more open options, e.g., by issuing skilled foreigners 10-year free visas to attract more talent and boost its economy (Leng, 2018).	Immigration control is a tool that is leveraged by governments to spur technological progress. The US has been selectively open to technical talent for decades.

Recommendations

The following are several general preliminary recommendations that can be made to governmental actors:

- **Increase efforts to collaborate with industry and academia:** The hope is that by sharing the same interests, governments, academia, and industry will strengthen connections among themselves. The UK Centre for Data Ethics & Innovation is a promising step taken by the UK government toward a proactive approach to innovation and data use in relation to AI. Governments may also consider investing in AI, focusing on fairness, accountability, and transparency. A specific area of interest may be to establish an AI verification agenda: given a strong military focus on AI, safety may be increased by requiring AI research to meet certain verification standards that support safety.

- **Prevent malicious use by non-state actors:** An example of an effort to prevent malicious use of technologies by terrorists, criminals, and other actors is the Campaign to Stop Killer Robots, supported by an

Open Letter signed by researchers in academia and industry, and by a video.

■ **Changes to immigration law:** Allowing Machine Learning PhDs who are foreign nationals to stay in their host country longer after they've earned their degree may help to foster long-term international fellowship and increasingly collaborative transnational communities.

Further research

The following considerations for creating effective national policy changes require further investigation in the context of AGI coordination:

■ **Incentivizing the incentivizer:** The creation and analysis of an incentive system has to include the institutions that are charged with producing that incentive system. If the incentive system being created does not sustain itself with regard to the incentives motivating the actors creating the incentive structure, the incentive structure is unlikely to be functional. This general warning about incentive design requires further research in regard to existing national and international policy efforts and future efforts in AI policy.

■ **Precedent cases**: Policymaking generally tends to be reactive, so researching available precedents for specific technologies can prove useful for future evaluation. For example, when the US Air Force set policy on the use of iPads in the US military prior to any relevant experience with using them, it had to determine threshold questions such as whether to treat iPads as phones or computers -- without any firsthand knowledge of the technology in question.

■ **Bridging communication gaps**: A general problem for fully informed US technology policy making concerns the physical gap between policy-shaping Washington, DC, and the technology-shaping Bay Area, which can translate into information gaps. Several promising projects might help bridge that gap, e.g., TechCongress, and Tech Foundry. Yet further research and action is required to translate those efforts into avenues for AGI policy development.

Military strategy and AGI Coordination

Framing effects

Discussions about Chinese and American military strategies may be influenced by psychological tendencies and biases, which can create risks in themselves. A tendency to exaggerate on both sides (e.g., with translations being slightly sensationalized by experts to gain reputation and media coverage), may result in a cycle of exaggerated claims feeding each other. This cycle could lead to negative self-fulfilling prophecies and increasingly adversarial and tense relations among different actors.

Risks

There are several military risks that could be significantly exacerbated by AI development. Two examples are:

■ **Military competition for AI**: Recent military developments suggest that competition for military

applications of AI is already underway between China and the US (Barnes, 2018). A specific focus of this competition is on the use of AI in military decision-making and in fighter planes, via research both on AI algorithms and building better neuromorphic supercomputers (Seffers, 2018). A detailed overview of China's military strategy can be found in China's Evolving Military Strategy (McReynolds, 2017). Even on a more general level, China has severely outspent the US government in AI investment, leading to strong trajectories for growth: "According to In-Q-Tel, an investment arm of the United States intelligence community, the U.S. government spent an estimated $1.2 billion on unclassified A.I. programs in 2016. The Chinese government, in its current five-year plan, has committed a hundred and fifty billion dollars to A.I (Larson, 2018)."

■ **Increasing risks from non-state actors**: While one coordination problem concerns big rivalrous state actors like China and the US, another risk concerns the ease with which non-state actors can build AI into autonomous vehicles, e.g., drones. The creation of autonomous aerial vehicles (UAVs) that the Campaign to Stop Killer Robots warns about may accelerate individuals' ability to cause global destruction and thereby shift the global power balance toward offense dominating over defense. The window for action on autonomous killer robots, including UAVs is rapidly closing. For potential paths to preventing their spread, see an Open Letter signed by researchers in academia and industry, and a video released by Future of Life Institute.

Complications with governance-based coordination efforts

General skepticism prevails about the chances of success for any effort to engage national actors in a conversation about decreased application of AI in the military. Strong incentives for militarization of AI are inevitable in the face of perceptions about potential AI militarization by other nations. For a detailed overview of AI's future impact on national defense capabilities, see AI And the Future of Defence (De Spiegeleire et al, 2018). Several factors cast doubt on the capacity of Western governments and institutions to be effective leaders in AGI coordination: similarities to the nuclear race (as discussed earlier in this report), where state actors arguably optimized for the wrong goals (see also Daniel Ellsberg's Doomsday Machine); the tendency for large institutions to be bureaucratic and slow-moving; and the poor track record of committees in designing and executing long-term, multi-step plans (Burja, 2018). An alternative strategy could focus on approaching civil society, academia and certain key stakeholders who are strategically situated to understand their counterparts in other nations. Individuals and their communities may become increasingly important to fostering coordination:

■ **The importance of individuals and their communities**: In contrast to governmental agencies, an interesting feature of empowered individuals is their relative freedom of action, suggesting that a focus on important individuals could be useful. There may be a window of opportunity for private actors to impact AGI safety trajectories, with coordination among key actors paving the way for public action. Best practices for AI could be codified, refined and evolved over time, before state actors enact and enforce regulation. Civil society and/or industry actors, such as research scientists, could create soft safety norms, which could gradually be crystallized into stronger decision-relevant iterative norms. Actors who do not respect these norms may be ostracized at first, before official enforcement mechanisms are overlaid on these norms. This strategy alone is unlikely to be sufficient in the case of AGI safety because exertion of social pressure often applied to ensure cooperation with norms, may be too weak to have the needed influence in this context. The fact that people usually play by social norms for low-stakes purposes is not satisfactory in the case of AGI. It is unclear how the dynamics of cooperation according to social norms play out when the stakes are high and there is no chance to learn from mistakes. It would be useful to analyze how norms form around use of

technologies that raise ethical concerns, and how to ensure they are not violated, for instance by analyzing analogous historial examples.

■ **The importance of individual actors within a community**: Another interesting dichotomy relates to the role of the individual within a community or company. Machine learning researchers' perceptions of important goals may be a potentially important lever for influencing important companies' strategies. If respected AI researchers refrain from joining companies that are perceived as bad actors or deviate from unsafe norms within a company, this push back could serve as strong signals for the rest of the research community to follow. It is unclear whether such protest strategies will be allowed or have a positive influence in non-Western countries. It would be useful to carefully identify conversational counterparts in academia, research, and civil society who can be engaged as regular interlocutors in informal conversations on important points of coordination in the future.

Recommendations

To optimize the role individuals can play in coordination of responsible AGI development vis a vis military strategies, detailed discussions among individual experts and cultural exchange among relevant actors on a personal level will be indispensable. Strategies that individuals may use to help foster a climate conducive to policy coordination include:

■ **Facilitating personal cultural exchange:** Conversations with government officials may be more useful when focused on intimate gatherings with specific individuals, rather than large media engagements, because the latter are often framed by adversarial narratives. Conversations surrounding major power imbalances need to be approached pragmatically, respecting a common base of reality and with careful consideration of context, framing, and mirroring effects. While the US may have concerns regarding the Chinese military spending on AI projects, within the US, DARPA is the largest government agency funding source for AI research, which allows for similar concerns on the Chinese side. A pragmatic approach that acknowledges the realistic concerns and needs of national actors is required to establish a common base reality for negotiations and cooperation. While military-to-military dialogue seems unrealistic, the nuclear arms race provides some precedent to inform such dialogues on a diplomatic level between the US and Russia, even though those dialogues were held at a stage where the proliferation of the relevant weapons systems already existed. Assuming good faith intentions extend beyond the U.S, facilitating cultural exchange by encouraging technical workshops and conferences in different countries will be productive. One conference conducted with this goal was the 2018 China US Tech Summit, hosted by The Future Society.

■ **Using available levers to influence important actors**: Existing policy levers may be used as vehicles for engaging key actors on AI policy. Tactics may include signalling potentially bad outcomes like reputation damage, supporting or proposing regulatory mechanisms, helping to implement new initiatives, and advocating for mutually beneficial relationships and coalitions.

■ **Cultural memes**: Given the precedent in the US government for popular movies influencing issue framing and politics and the high affinity for movies in China, one might look to the entertainment industry as a possible platform for advocating in favor of Asilomar-esque principles of safety. Some historical examples of culture-shaping movies in the US include the movies War Games, The Day After, and Reagan. In the long run it might be possible to influence military practices by creating a generally more collaborative culture. However, even if the public was engaged in a collaborative AGI narrative, the extent to which this would

suffice to lessen AI military arms race dynamics among nations is unclear. If the potential for a military arms race cannot sufficiently be influenced by civilians, it remains a catastrophic risk factor that needs to be addressed.

Further research

An interesting possibility requiring further investigation is whether a more positive 'race' or contest could be encouraged, not toward greater development of AI capabilities but to predictability, stability, and security of AI. If the narrative of the desired AI outcomes could be shaped toward making inherently safe dynamics more attractive to actors, such as by publicly promoting important safety initiatives, race dynamics may be steerable toward safety over capability.

2b. AGI Coordination Scenarios: Major Private Actors

Private actors

Framing effects

A variety of potential framing effects may be relevant to the coordination of large private actors:

- **Vocabulary**: As AI safety concerns are still fairly nascent in China, a careful approach to creating positive AI safety narratives is essential. The western framing of "AI safety" may be difficult to apply in the Asian context because "safety" could be interpreted as being counterproductive to "efficiency"—a core value in the Chinese context, as illustrated by China's swift deployment of self-driving cars. Other narrative differences between the US and China pertain to the public perception of AI. For example, US consumers may be influenced by dystopian depictions of AI in Western movies, while Asian consumers may tend to see AI as "cute."

- **Mirror-imaging effects**: One foundational discussion point on AGI coordination concerns effective avenues to propagate safe AI strategies. One may consider whether top-level propagation via organizational influence on the one hand, or active cooperation on an individual research level on the other, is more effective in spreading safety standards. The answer to this question may vary significantly by countries. Recently in the US, there have been several signs of cooperation by private actors (see below) whereas in the context of China, it is questionable whether organizational statements of commitment are credible and effective cooperation signals; arguably those signals would be more impactful if coming from the political leadership. Investigating signals for cooperation that are effective in different political contexts would be valuable.

Comparing different types of actors

The Chinese government has, at least on the surface, a stronger influence on many organizational operations than is assumed for the US. Those de facto "private-public partnerships" are often misconstrued as pure monolithic government strategies. A breakout of the actors involved (see below) can help bring some clarity, notwithstanding limited knowledge of the space of actors within China generally, and limited knowledge of the exact goals and focus of the organizations listed, whether those include AGI or AI alone:

2b. AGI Coordination Scenarios: Major Private Actors

	Industry	Academia	Governmental
Examples	Western industry efforts in AI include: DeepMind, OpenAI, Google Brain, FAIR and other projects, e.g. Vicarious, Palantir, Amazon, Microsoft, IBM, and GoodAI. Chinese efforts include Baidu, Alibaba, and Tencent's gaming and security research.	Western academic efforts in AI include Montreal's MILA, Berkeley's BAIR, IDSIA, the Stanford AI Lab, Cambridge's CBL, and European ELLIS. Although little information is available about the Asian landscape, efforts may include those of the Chinese military, Tencent, Baidu, or academia, such as Tsinghua University or the Chinese Academy of Sciences.	Government-focused work includes DARPA's AI work in the US, the ITU's Alliance for Good, International Organization on Standardization group on AI, for the first time co-chaired by China and the US, and AI Industry Alliance (AIIA) launched in 2018, with the goal of ensuring that China's industry adjusts smoothly to technological acceleration from AI breakthroughs.
Progress	Industry may spend most of its energy on profitable short-term AI tools (e.g., TensorFlow) rather than on the basic research required for paradigm-shifting tool building. However, DeepMind can be interpreted as a case in point for an attempt at a long-term 'moonshot' by Google, similar to OpenAI. Deepmind and OpenAI have also been spending considerable resources to accomplish feats that capture the public's imagination (e.g., AlphaGo, Dota) while universities and other corporate research groups are less focused on public perceptions. OpenAI and corporate research groups such as Google Brain and DeepMind have done a lot of research that takes advantage of their vast amounts of computational resources.	Students at universities and smaller industrial labs are generally restricted in the amount of computation they can employ. At some relatively well-funded labs, students can use a moderate amount using Google Cloud and AWS but they nevertheless do not have the 'big compute' resources that the top industry labs have.	It can be argued that DARPA's work is more short-term than its reputation suggests. A potential exception to this generalization is DARPA's Active Interpretation of Disparate Alternatives (AIDA), which has as a goal to "develop a multi-hypothesis semantic engine that generates explicit alternative interpretations of events, situations and trends from a variety of unstructured sources, for use in noisy, conflicting, and potentially deceptive information environments."
Collaborativeness	Google Brain, OpenAI, and DeepMind have published several papers together, The Partnership on AI recently launched, uniting multiple large industry actors under one umbrella organization. Otherwise there is not much collaboration among different industry actors.	There is a lot of intermarriage between industry and academia with many professors especially at top institutions taking on industry roles while keeping their academic appointments and having their students intern at corporate labs. Finally, there is a recent surge of interest in startups in some academic labs (e.g., at Stanford) among both students and faculty (which is far less common among researchers in industry labs).	China's AI Industry Alliance is led by the China Center for Information Industry Development, and backed by 240 Chinese technology companies, including giants like Intel China, iFlytek Co Ltd, JD.com, SAP China and Ecovacs Robotics Co. The alliance set goals of incubating 50 AI-enabled products and 40 firms, launching 20 pilot projects, and setting up a general technology platform in the next three years (Shuiyu, 2018).

Recommendations

Fostering a collaborative research culture, especially in relation to safety, is an important starting point for coordination. The following are a few examples of organizations that have led the way by committing to safety and signalling cooperation via public statements, value statements, or collaborative action with other organizations:

- Collaborative research by OpenAI and DeepMind, specifically on Learning Through Human Feedback (Legg, Leike, Martic, 2017).
- Partnership on AI's value tenets, which commits to an open dialogue on the ethical, social, economic and legal implications of AI, and fostering a culture of cooperation, trust, and openness among AI scientists and engineers (PAI, 2018).
- The OpenAI Charter includes a Chinese translation and states that its mission is for AGI to benefit all of humanity for instance by focusing on broadly distributed benefits, long-term safety, technical leadership, and cooperative orientation. Specific principles send clear signs for cooperation, as indicated by these statements: "We are concerned about late-stage AGI development becoming a competitive race without time for adequate safety precautions. Therefore, if a value-aligned, safety-conscious project comes close to building AGI before we do, we commit to stop competing with and start assisting this project," and "We will actively cooperate with other research and policy institutions; we seek to create a global community working together to address AGI's global challenges" (OpenAI, 2018).

Public leadership of this nature by respected organizations is a valuable sign of hope for broader cooperation, especially because the public nature of their statements allows for accountability. It would be useful if other organizations issued similarly credible, cooperative statements to signal their commitment to avoiding arms races. Such signals would be especially valuable coming from organizations that strongly focus on capabilities research in addition to, or instead of, safety research.

Further research

The value of information is a topic that is becoming increasingly important especially in relation to coordination among private actors in AI. Further research is required to distinguish the value and potential risks of information in different contexts. Several publications discuss the role of information pertaining to AI, especially to AI capability, e.g., Strategic Implications of Openness in AI Development (Bostrom, 2017), and Racing to the Precipice: A Model of AI Development (Armstrong, Bostrom, Schulman, 2017). It suggests that an AI race could develop if AI technology is seen as overwhelmingly powerful, with great first mover advantages and little possibility for the losing "side" to catch up. The degree of risk is calibrated not so much on whether this is true, but whether this is believed to be so by the teams involved. In that situation, there is a strong incentive to skimp on safety precautions, and try to rush to achieve the AI goal. This is dangerous, because if the AI is as powerful as feared, skimping on safety can be disastrous even for the "successful" team. Several rather obvious measures can help to reduce the pressure to race: fewer teams, better coordination between teams, and shared values are all helpful.

On the other hand, the value of information can be positive or negative, in theory. If one team is very far ahead of all the others, sharing this knowledge can reduce risks, especially if the leading team knows they can afford to be careful. If all the teams are roughly at the same level, then detailed sharing of capacity knowledge could be disastrous, as each team sees an incentive to continually skimp a bit more on safety, to get ahead of its rivals. Although these results have theoretical backing, in practice, the strongest effect of information availability is likely to be that teams

that share more information can trust each other more, improving coordination. It is hoped that mechanisms and agreements will be put into place early to facilitate substantive agreements and coordination between developers, and reduce the risks of AI arms races (Armstrong, 2016).

A few proposals for handling information on AI in novel ways require further investigation:

- **Rethinking openness**: A recent report published by FHI on The Malicious Use of Artificial Intelligence highlights the dual use of AI and ML approaches and proposes to reimagine norms and institutions around the openness of research. Such innovation could occur via pre-publication risk assessment in technical areas of special concern, central access licensing models, and sharing regimes that favor safety and security (Brundage, 2018).
- **Selective information**: Additionally, the value of information is not just a binary question of more or less information, but requires more precise investigation of what type of information is adequate and necessary to inform which action.

Incentivizing safety

Lack of tractability of safety research

Arguably safety, fairness, accountability, and transparency are integral aspects of well-designed AI systems because "an AI is misaligned whenever it chooses behaviors based on a reward function that is different from the true welfare of relevant humans" (Hadfield-Menell, Hadfield, 2018). Despite the foundational importance of safety, safety research tends to lag behind capability research because it is harder to track progress on what it means to be aligned with human values than concurrent progress on capability. While there is some fairness research that aims at finding a fairness metric to optimize for, most metrics are still fuzzy and uncertain compared to more tractable capability problems. The relative intractability of safety problems may disincentivize safety research in comparison to capability research within organizations and in journal submissions:

- **Within organizations**: : While many AI safety organizations have an explicit focus on safety, much of the published work is focused on capability-related research, rather than tackling specific safety problems. This emphasis is partly owing to the practical reality that researchers have incentives to make tangible progress on open research problems and publish clear results to compete for research positions and grants. Given that it is more difficult to make tangible progress on less tractable safety problems, there is a disincentive for researchers to work toward safety.
- **Journal and conference submissions**: Most prestigious AI journals and AI conferences were historically not focused on safety. Given that journal publications and conference presentations are part of the means by which AI researchers signal their skills and reputation, this omission may lead to additional disincentivization of safety research. However, with AI and AI safety organizations caring less and less about traditional academic signals and benchmarks in their hires, this problem may be alleviated in the future.

Recommendations

Potential solutions incentivizing AI safety research within organizations and via journals and conferences involve concretizing safety research agendas to produce more tractable outcomes, and highlighting the security aspects

of AI safety. More specific ideas include:

- **Increasing incentives for work on safety within organizations:** Previously a major bottleneck in incentivizing safety research was skeptical views among senior researchers. Many of them now acknowledge safety as an important challenge, thereby lessening reputation problems. To further incentivize safety work, those senior researchers could establish concrete safety research agendas. Promising safety research overviews (Mallah, 2017) and research guidelines have been published by Paul Christiano, Dario Amodei and others, Stuart Russell and others and MIRI. Some of them could be made more concrete so that they are attractive to and manageable by more junior researchers.

- **Increasing incentives for safety work set by journals and conferences:** A strategy that is especially applicable to increasing the representation of safety research in journals and conferences is to rebrand safety research as matter of security. By framing "AI safety" concerns as "AI security" concerns, the innate connection between safety/security on one hand and capability on the other hand is made clear, thereby demonstrating that real-world application of insecure AI is untenable.

Further research

Research to concretize the safety research agendas listed above would be highly valuable because it would incentivize junior researchers to start working on safety and increasingly move those concerns into the Overton Window. In addition to concretizing safety problems, research into partial solutions to existing safety problems may also encourage traction for safety concerns by showing that some initial understanding and progress on the issue is possible, making it more difficult for skeptics to reject those issues out of hand.

3. Technological Factors for AGI Coordination: Challenges and Potentials

Cybersecurity

Framing effects

Discussions of Chinese and US approaches to cyber-technologies are likely to be influenced by a number of framing effects. From a Chinese perspective for instance, the American framing of the discussion of individual rights could be challenging because many of the cyber-rights violations that China is criticized for are arguably performed by the US also via domestic surveillance by the NSA, or via the US 5 EYES partners, who surveil on the US' behalf. Thus, to avoid cutting negotiations short, an attitude of humility and understanding rather than blame is appropriate at the negotiation table.

Cybersecurity as a factor for Catastrophic and Existential Risk

The current state of computer security may arguably already present a catastrophic risk to the world.
Many computer systems that are currently being deployed and exploited are not only insecure, but insecurable. Grave vulnerabilities exist in hardware, software, and operating systems. While both software and operating systems are hard to secure, hardware is one of the most difficult parts to secure as it requires a secure chain of custody, trusted manufacturing and successful addressing of other hard-to-control factors. Thus, even if one used secure software, operated on a secure operating system, the hardware that those are running on is likely to be fundamentally insecure. Given how heavily China-reliant the current hardware supply chain for US-based manufacturing is, exploits could currently be inserted in most hardware. Because vulnerabilities are networked, local vulnerabilities compound into regional vulnerabilities, which compound into international vulnerabilities, increasing the risk of large-scale attacks (Peterson, Miller, Duettmann, 2017).
According to the Snowden revelations of 2013, the capability of deep surveillance and infiltration into nearly all modern hardware is proven and rampant. Every stage in global supply chains represents a vector of possible compromise. State actors have made a concerted effort (such as Project Bullrun) to undermine global cryptography

standards upon which much of the global economy and security is based. Even if end-to-end encryption is employed and remains resilient to attack, the underlying hardware used to enable that encryption is often trivially infiltrated. Large incentive/coercion programs exist whereby state actors compromise and inspect the software and private data of major technology companies, including source code access. For example, Microsoft has shared its source code with the Chinese government since 2003; flaws found within that code enabled numerous cyber attacks at dozens of companies and countries that used Microsoft software. The modern computing landscape is built on fundamentally insecurable infrastructure and software.

At the same time, more and more of society is reliant upon digital infrastructure with little thought given to redundancy. AI, even in narrow forms, multiplies computer security risks because it can be employed to more easily exploit the weaknesses in existing social and technical systems. Any capability one state actor develops is likely to be replicated by others in short order, and even non-state actors can represent an advanced and persistent threat. Several states have obtained cyber weapons that can cause trillions of dollars in damage, and potentially are as destructive as nuclear weapons. However, unlike nuclear weapons, there are few deterrents to cyber weapons since attribution of a cyberattack can be difficult or impossible to ascertain and escalation has fewer consequences for the attacker.

The current level of risk associated with cyber insecurity is severe and endemic, and poses a Catastrophic Risk in itself, while also potentially exacerbating other risks with potential catastrophic or existential consequences:

- **The electric grid**: An example of a serious, potentially catastrophic attack that is possible with current capabilities involves the electric grid. The U.S. electric grid for instance, "is vulnerable today to cyber attack with damage estimates by Lloyd's ranging up to $1 trillion" (Rashid, 2015). Damage to the electric grid via cyber attack can include physical as well as software damage, and would take months or arguably, years to repair, leaving an entire multi-state region without power. Lloyd's, as an insurance company, focused on estimating financial damages rather than fatalities. While plans have been made at the federal level in the U.S., they were prepared under a previous administration, and it is as yet unclear whether these or similar plans will be carried out (Peterson, Miller, Duettmann, 2017).

- **Totalitarian governance**: An example of a potential existential risk that may be exacerbated by current trajectories of cyber-technologies concerns the creation of a totalitarian agency that could control the world. In addition to its citizen score experiment, China is increasingly tightening its Great Firewall by declaring unauthorized VPN services illegal and forcing both local and foreign companies and individuals to use only government-approved software to access the global internet (Ye, 2018). While the US does not prohibit VPNs or regulate Internet access, the Snowden revelations established that the NSA's surveillance and infiltration techniques were deep and pervasive years ago, and there is little reason to believe that they have become less powerful since. While in themselves problematic, such developments make future abuse of power more likely and harmful.

Recommendations

To prevent risks from AI built on top of insecurable computer foundations, and to prevent other risks arising from the lack of computer security, relevant computer systems should be moved to new security architectures. In theory,

there are several potential avenues to improved global security of computer systems:

- **Defense in Depth**: This is a critical infrastructure defense strategy currently employed by militaries, that assumes that some systems will be compromised, and emphasizes learning from compromises, and retaining security of the most important, deepest network. This approach has the advantage of allowing the defender to learn about the attacker because the different firewalls that the attacker breaches allow for intelligence to be captured. However, it is unclear that this approach will survive in a world where AI is directed towards malicious exploitation:

> "This situation is only survivable because the attacks that nation-states are developing are probably much less sophisticated than the attacks that the most advanced organizations could be engaging in by making better use of bleeding-edge early technologies combined with static analysis technologies. For instance, the strategies that are known from the Snowden revelations include gathering Zero-Day Attacks, i.e., entities wanting to take over others' computers accumulate Zero-Day Attacks to prepare for a future day when that entity will use them against those target computers owned by others (Wikileaks, 2013). However, rather than gathering known Zero-Day Attacks, one can imagine software that is able to analyze the software being attacked and find entirely new, previously unknown Zero-Day Attacks. Having the best state-of-the-art software for discovering vulnerabilities built into the deployed attacking system would enable the system to discover vulnerabilities and exploit them while it is in active contact with the target, rather than just launching built-in attacks against previously known vulnerabilities. This level of attack software is one that the currently entrenched architectures are not going to survive, and it is likely to precede AGI." (Peterson, Miller, Duettmann, 2017).

- **Technical solutions**: Technical solutions to security vulnerabilities are available and practicable. For example, the seL4 microkernel is our best case of an operating system that seems to be secure, due to its formal proof of end-to-end security and its track record of having withstood a Red Team Attack (a full-scope, multilayered attack simulation), which no other software has withstood (Brundage, 2018). One hopeful development is increased funding for seL4 by the U.S. Department of Defense. Nevertheless, its security rests on some counterfactual assumptions, such as that the formal model of the underlying hardware is accurate.

- **Responsible disclosure:** Responsible disclosure with timelines is a defense suggestion for finding existing vulnerabilities in the wild, and disclosing those privately to the affected organizations, while giving them a certain period to resolve the vulnerability before making them public. Responsible disclosure is already practiced as the norm within the cybersecurity community (Brundage, 2018). Given the NSA's dual mandate of civil defense and offense to secure defense, the NSA could function as vulnerability collector and disclosing apparatus. A 12-year timeline could be set for the NSA zero-day vulnerabilities that are held by the government (Duettmann, 2017).

- **The blockchain ecosystem**: A strategy for defense, discussed in more detail in the next section of this report, concerns the gradual move to a blockchain-based ecosystem. Currently, chances of deployment of any secure infrastructure are low because a multi-trillion dollar ecosystem is already built on the current insecurable foundations, and it is very difficult to achieve adoption of something that requires the entire ecosystem to be rebuilt from scratch. Researchers have been exploring strategies to bridge from current systems to new secure ones, in a process analogous to what in a biological context is known as "genetic takeover" (Peterson, Miller, Duettmann, 2017).

Further research

Currently, none of the above avenues to increased security are being explored thoroughly. Progress on offensive cybersecurity is growing rapidly, and it is possible that novel technological developments can exacerbate insecurities in the absence of an explicit focus on strengthening defense. One controversial example is the future effects of quantum computing on cybersecurity—an area in which China is making rapid strides. What quantum computing does and does not mean for encryption is a topic that requires further research. However, it is a noncontroversial expectation that 'business as usual' in cybersecurity research and deployment will likely lead to cybersecurity offense dominating defense over the long run, leading to increased catastrophic risk quite apart from its relationship to AGI.

Blockchain & Cryptocurrency

The blockchain ecosystem and AI safety

While there is some skepticism about blockchain's potential for AI coordination, there are a number of potentially beneficial use cases of blockchain technology and its emerging ecosystem for AI safety:

- **Blockchain ecosystem as role-model for secure computation:** Both Bitcoin and Ethereum are evolving in an ecosystem that is already under very hostile attack pressures because projects often create the practical equivalent of a multimillion dollar cryptocurrency 'bug bounty.' When insecurity leads to losses, the actors have no other recourse to compensate them than getting a majority of users to agree to a major rollback. Systems that are not bulletproof will be killed early and visibly, and therefore these ecosystems remain populated only by apparently bulletproof systems. The bulletproof security of these systems is an essential part of their value proposition. Such projects are evolving with a degree of adversarial testing that can create the seeds for a system that can survive a magnitude of cyberattack that would destroy conventional software. If this type of secure system matures sufficiently before the world is subject to that type of cyberattacks, then a successful 'genetic takeover' scenario might be achieved (Peterson, Miller, Duettmann, 2017).

- **Computational law for artificial agents**: To the extent that blockchain enables the making of rules that not only humans could be following but that machines could be encoded to follow, there is an opportunity to craft laws that will affect how AIs and humans operate and interact in the world (Peterson, Miller, Duettmann, 2017).

- **Importance of cross-jurisdictional coordination**: Given that blockchain is cross-jurisdictional in nature, the extent to which its computational law can be enforced remains to be seen. However, computational law also provides a unique framing for solving cross-jurisdictional coordination problems. Often certain legal policy framings haven't been concretely articulated into laws, leaving a liminal blank space, which is regulated by different policymakers differently. Many of the arrangements advocated in the blockchain-space are not illegal, but are genuinely novel economic phenomena, comparable to the way how in the early days of the internet it was unclear whether purchases made by credit cards were legal, or how Supreme Court decisions struck down regulations of railroads because those regulations were written prior to sufficient experience

with the railroad system. Similarly, the blockchain ecosystem is creating a new law-governed system that can succeed at creating beneficial rule-based interactions. This process should follow a path that mirrors the trans-jurisdictional nature of the internet and could open up novel legal avenues for global coordination.

■ **Proof of location:** The Sybil problem (i.e., individual actors acting like a great number of actors so that their influence becomes a majority of the voting power determining the future of the system) remains a problem within the blockchain ecosystem, especially concerning high-stakes applications of the blockchain. Besides Proof of Stake, a potential solution includes Proof of Location, an economic attestation that something is at a particular location at a particular time, which could be used to incentivize geographic distribution, e.g., by increasing block rewards in a certain location. POL could be used to prove geographic diversity, thereby encouraging jurisdictional diversity and may be a useful tool to facilitate global private coordination efforts that circumvent national jurisdictional boundaries.

■ **Prediction-markets as incentive markets:** Using blockchain for the creation of large-scale AI safety prizes to spur research may not be necessary because those prizes typically have no credibility problems which could be resolved by deploying blockchain. However, an alternative incentive-setting mechanism, which would be enhanced by blockchain technology is the creation of incentive markets to foster research on certain AI safety domains. Gnosis, a live-audited smart contract for prediction and incentive markets, is currently working with different jurisdictions to launch commercially. The advantage of incentive markets is that the incentive-setting can occur in an anonymous, cross-jurisdictional and incremental way, aimed at a top-level goal with conditional markets below. Those markets do not have to be limited to a few large bounties, but researchers who make incremental breakthroughs in safety work could buy shares in the positive outcome of related research, creating researcher swarms working toward safety, rather than letting only one team claim a big prize. The open nature of incentive markets for safe, public-interest AGI is a promising approach to open and democratic development of AGI safety. The potential negative usages of those markets, such as incentivizing illegal actions, need to be investigated and weighed against their potential benefits.

■ **Smart contracts for cooperation**: The potential use of smart contracts for greater cooperation among powerful entities encounters certain problems. For instance, it may be difficult or impossible to draft complete contracts, and connect those contracts to real-world repercussions. However, even if the ability of blockchains to create enforceable obligations between nations is complicated, it may serve as a transparency and public record auditability mechanism, by recording all actions taken within a contract on tamper-evident logs.

Recommendations

Although the role of blockchain for AI safety requires further research, several preliminary recommendations can be made to both individuals in the blockchain space and to governments approaching the blockchain space:

■ **The importance of individuals in the cryptocurrency community**: While much of the AGI coordination discussion focuses on governments or industry actors, the cryptocurrency and blockchain narrative allows for an investigation of the role of the individual. There is now an emergent community around cryptocurrency whose ideology centers very much around privacy, data-ownership, and anti-surveillance. This cryptocurrency community also has an interest in AI being pursued safely and securely. To the extent

that individual actors in this space now have significant resources at their disposal, which may dwarf even traditional funding sources, there is a clear potential for them to support AI safety efforts that are goal-aligned with their own existing interests. Multiple non-profit organizations working on AI safety already accept cryptocurrency (e.g., EFF, Foresight Institute, MIRI).

- **Incentivizing reasonable governance of blockchain:** Given the infancy of the blockchain space and the lack of expertise on the government side, there is a limited window of opportunity to encourage sensible regulation, developed jointly by governments and actors in the cryptocurrency space. Specifically, participants in the AGI strategy meeting discussed a project that would gather important individuals in the blockchain space to reach out to governments via a white paper that lays out a detailed analysis of how they could leverage the smart contract ecosystem to benefit individual sectors of the economy. Such a white paper could greatly benefit different sectors, including education, employment and healthcare, and would serve as strong signal for cooperation from the crypto-community toward governments, that could help avoid pre-emptive hyper-regulation of the blockchain ecosystem. Simultaneously, by applying sovereignty-giving blockchain tools to deliver improved services in relevant sectors, governments have the chance to rebuild decaying trust in institutions and compete with other countries around citizen-aligned government. Please contact Foresight Institute with interest in this project.

Further research

Currently, there is scant research focus on the role of blockchain in the Western and Asian context with respect to AI. A deeper understanding of the space would be useful for investigating potential effects on AI safety. A few key factors that may serve as starting points for further research include:

- **Blockchain in China:** China generally recognizes the importance of blockchain technology, encouraging domestic development to compete with international development. The People's Daily, which gives an insight into the opinions of the Chinese Communist Party, published an article on Three Questions to Blockchain, stating that "the mainstream blockchain technology platforms all originated from abroad. The domestic blockchain technology service providers should patiently start from the ground floor to make technologies independent and controllable, and strive to lead global blockchain technology development." While there are several cryptocurrencies originating in China, most are developed by actors that are not CCP members, with the exception of the Matrix AI Network, which fuses AI and blockchain. The Matrix Network has just signed a strategic cooperation agreement with the state-owned $900 billion Belt and Road Development Centre to become the only blockchain partner of the center. For more information on the Matrix AI Network's scale, see this article on "The Biggest Crypto Partnership the World Has And Will Ever See".

- **Cryptocurrency in China:** Scaling computer factories in Shenzhen and the increasing East/West mining divide lead to worries of mining pools being dominated by Asian actors. Allegedly, a recent CPC crackdown on illicit local government mining revealed that local governments diverted hydropower to miners, being paid in Bitcoin. Arguably, this problem is not specific to Asian mining pools, but the whole blockchain system is premised on the Satoshi solution to the Sybil problem, which does not sufficiently disincentivize scaling.

Conclusions

Each section in this report contains individual conclusions that are included in the Executive Summary. Finally, a few factors that merit especially high priority for further investigation, because they are relevant to all coordination scenarios on AGI, are:

1) **Framing effects:** Framing of AI narratives, biases, and the different considerations that drive decision-making in different cultural and social contexts were independently discussed for each coordination scenario. Framing effects on AI safety are likely present in this report itself, and careful outreach among different actors is necessary for them to acquire a reciprocal understanding of each other.

2) **Cybersecurity**: Current cyber-insecurities, paired with current and future offensive capabilities that can be exploited by state and by non-state actors alike, make cyber-insecurity a risk with potentially catastrophic consequences in itself, in addition to its exacerbating consequences for AGI risk. Any actors' coordination effort that does not take its vulnerability to potential cyberattacks into account cannot realistically succeed.

3) **The importance of individual actors**: Individual actors who can have a great impact on coordination include decision-makers in AGI organizations and individual researchers within the AI community, but also external actors, for instance within the cryptocurrency community, who may be goal-aligned and situated in a unique position to support important AGI safety efforts.

Framing effects, cybersecurity, and the importance of individual actors are relevant not only for AGI coordination among great powers but are likely to be important in solving a variety of distinct future coordination problems, such as those arising from potential biotechnology weapons. While most claims in this report are AGI-specific, many of the recommendations on coordination may also provide useful starting points for creating an overall policy framework that is antifragile in the face of novel risks. Working toward a world in which we are better prepared to face a variety of risks is a strategy that is discussed in Existential Risks and Existential Hope: Definitions (Ord, Cotton Barratt, 2017), which led Foresight Institute to launch Existentialhope.com, a collaborative knowledge graph and global progress tracker of key areas important for the future of life. The section on AI & Cyberspace contains further suggested reading on topics relevant to AI coordination and AI safety more generally.

References

Afanasjeva, O., Feyereisl, J., Havdra, M. 2017. "Avoiding the Precipice: Race Avoidance in the Development of AI." AI Roadmap Institute.
https://medium.com/ai-roadmap-institute/avoiding-the-precipice-db720a805190

Amodei, D. Olah, C. Steinhardt, J. Christiano, P. Schulman, J. Mane, D. Concrete Problems in AI Safety". ARXIV.
https://arxiv.org/pdf/1606.06565.pdf%20http://arxiv.org/abs/1606.06565.pdf

Armstrong, S., Bostrom, N., Schulman, C. 2016. "Racing to the Precipice: A Model of Artificial Intelligence Development". AI & Society.
https://link.springer.com/article/10.1007/s00146-015-0590-y

Barnes, J., Chin, J. 2018. "The New Arms Race in AI". Wall Street Journal.
https://www.wsj.com/articles/the-new-arms-race-in-ai-1520009261

Baum. S. 2018. "Superintelligence Skepticism as a Political Tool". Information, 2018, vol. 9, in press

Baum, S., Neufville, R., Barrett, A. 2018. "A model for the probability of nuclear war." Global Catastrophic Risk Institute Working Paper 18-1.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3137081

Baum, S., Barrett, A. 2018. "A model for the impacts of nuclear war." Global Catastrophic Risk Institute Working Paper 18-2.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3155983

Baum, S. 2017. "A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy." Global Catastrophic Risk Institute Working Paper.
https://poseidon01.ssrn.com/delivery.php?ID=188070086031106064070091006092107 1

References

100040110910520610610100861201260260680860070971190370111250631160000981010250730821000281040200750070330720060901150861130780251030060650340670140910720641210700830921181080981211030001240111160911070711021140200731116&EXT=pdf

Bensinger, R. 2018. "MIRI's 2018 Research Plans and Predictions." MIRI.
https://intelligence.org/2018/03/31/2018-research-plans/

Bensinger, Rob. 2016. "White House Submission and Report on AI Safety." MIRI
https://intelligence.org/2016/10/20/white-house-submissions-and-report-on-ai-safety/

Bostrom, N., Sandberg, A., Douglas, T. 2013. "The Unilateralist's Curse: The Case For a Principle Of Conformity". Future of Humanity Institute.
https://nickbostrom.com/papers/unilateralist.pdf

Bostrom, N., Dafoe, A., Flynn, C. 2017. "Policy Desiderata in the Development of Superintelligent AI." Future of Humanity Institute.
https://nickbostrom.com/papers/aipolicy.pdf

Bostrom, N. 2017. "Strategic Implications of Openness in AI Development." Global Policy.
https://nickbostrom.com/papers/openness.pdf

Burja, S. 2018. "Competition for Power." Medium.
https://medium.com/@samo.burja/competition-for-power-8da516d0b2b3

Burja, S. 2018. "Borrowed vs. Owned Power." Medium.
https://medium.com/@samo.burja/borrowed-versus-owned-power-a8334fbad1cd

Brundage, M. & Avin, S. et al. 2018. "The Malicious Use of AI." Future of Humanity Institute.
https://img1.wsimg.com/blobby/go/3d82daa4-97fe-4096-9c6b-376b92c619de/downloads/1c6q2kc4v_50335.pdf

Cairns-Smith, A.G., 1982. "Genetic Takeover And The Mineral Origins Of Life". Wiley.

Calo, R. 2015. "Robotics and the Lessons of Cyberlaw," California Law Review, Vol. 103, No. 3, pp. 513-63.

Carbon Black, 2017. "Beyond the Hype: Security Experts Weigh in on Artificial Intelligence, Machine Learning, and NonMalware Attacks."
https://www.carbonblack.com/2017/03/28/beyond-hype-security-experts-weigh-artificial-intelligencemachine-learning-non-malware-attacks/

Carlini, N., Mishra, P., Vaidya, T., Zhang, Y., Sherr, M., Shields, C., Wagner, D., and Zhou, W. 2016. "Hidden Voice Commands." 25th USENIX Security Symposium.

References

Cave, S., Ó hÉigeartaigh, S. 2018. "An AI Race for Strategic Advantage: Rhetoric and Risks." AIES Conference.
http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_163.pdf

Christiano, P. 2017. "Directions and Desiderata for AI Alignment." AI Alignment.
https://ai-alignment.com/directions-and-desiderata-for-ai-control-b60fca0da8f4

Cillizza, C. 2014. "Yes, Obama is right. The 113th Congress will be the least productive in history." Washington Post.
https://www.washingtonpost.com/news/the-fix/wp/2014/04/10/president-obama-said-the-113th-congress-is-the-least-productive-ever-is-he-right/?noredirect=on&utm_term=.7e2df8e5740f

Dennis, J., Van Horn, E. 1966."Programming semantics for multiprogrammed computations" Communications of the ACM.

Desilver, D. 2018. "Congress has long struggled to pass spending bills on time." Pew Research Center.
http://www.pewresearch.org/fact-tank/2018/01/16/congress-has-long-struggled-to-pass-spending-bills-on-time/

Ding, Jeffrey. 2018. "Deciphering China's AI Dream." Future of Humanity Institute.
https://www.fhi.ox.ac.uk/wp-content/uploads/Deciphering_Chinas_AI-Dream.pdf

Decker, S. 2018. "Forget The Trade War: China Wants to Win at Quantum Computing." Bloomberg.
https://www.bloomberg.com/news/articles/2018-04-08/forget-the-trade-war-china-wants-to-win-the-computing-arms-race

Duettmann, A. 2018. "AI Safety: State Of The Art." SXSW.
https://www.youtube.com/watch?v=Lg1FAtfSheo

Duettmann, A. 2017. "AGI Timelines & Policy White paper." Foresight Institute.
https://foresight.org/publications/AGI-Timeframes&PolicyWhitePaper.pdf

Duettmann, A. 2018. "AGI Safety: Overview & Definitions," Foresight Institute AGI & Corporations Seminar at the Internet Archive.
https://www.youtube.com/watch?v=l10_IUwB1-I

Eckersley, P. Cohn, C. 2018. "Google Should Not Help the US Military Build Unaccountable AI Systems." Electronic Frontier Foundation.
https://www.eff.org/deeplinks/2018/04/should-google-really-be-helping-us-military-build-ai-systems

Eckersley, P., Gillula, J., Williams, J. 2017. "EFF's submission to the House of Lords Select Committee on Artificial Intelligence (AI) request for comments." Electronic Frontier Foundation.
https://www.eff.org/document/eff-submission-house-lords-select-committee-ai
Eckersley, P. Nasser, Y. 2018. "Measuring the Progress of AI Research". Electronic Frontier Foundation.
https://www.eff.org/ai/metrics

References

Ellsberg, D. 2018. *The Doomsday Machine: Confessions of a Nuclear War Planner.* https://www.amazon.com/Doomsday-Machine-Confessions-Nuclear-Planner/dp/1608196704

Executive Office of the President. 2016. "National Electric Grid Security and Action Plan." https://www.whitehouse.gov/sites/whitehouse.gov/files/images/National_Electric_Grid_Action_Pl an_06Dec2016.pdf

Fisher, Kathleen. 2014. "Using formal methods to enable more secure vehicles: DARPA's HACMS program." Proceedings of the 19th ACM SIGPLAN international conference on Functional programming.

Future of Life Institute. 2017. "Asilomar AI Principles". Future of Life Institute. https://futureoflife.org/ai-principles/

Future of Life Institute. 2018. National And International AI Strategies. Future of Life Institute. https://futureoflife.org/national-international-ai-strategies/?cn-reloaded=1

Gallagher, R. 2018. "Google Plans To Launch Censored Search Engine in China, Leaked Document Says." The Intercept. https://theintercept.com/2018/08/01/google-china-search-engine-censorship/

Guan, M. Y. 2018. "Regulating AI in the era of big tech." The Gradient. https://thegradient.pub/regulating-ai-in-the-era-of-big-tech/

Hadfield-Mendell, D., Hadfield, G. 2018. "Incomplete Contracting And AI Alignment." https://arxiv.org/abs/1804.04268

Larson, C. 2018. "China's Massive Investment in AI Has an Insidious Downside" Sciencemag. http://www.sciencemag.org/news/2018/02/china-s-massive-investment-artificial-intelligence-has-insidious-downside

Legg, S., Leike, J., Martic, M. Learning Through Human Feedback. DeepMind. https://deepmind.com/blog/learning-through-human-feedback/

Leng, S. 2018. "Could you be a "high-end" foreigner? China offering 10-year free Visa to top talent," South China Morning Post. https://www.scmp.com/news/china/economy/article/2126882/could-you-be-high-end-foreigner-china-offers-10-year-free-visa

Lyons, T., Felten, E. 2016. "The White House Report on The Future of Artificial Intelligence." The White House. https://obamawhitehouse.archives.gov/blog/2016/10/12/administrations-report-future-artificial-intelligence

Liu, H., Cedervall Lauta, K., Maas, M. 2018. "Governing Boring Apocalypses: A New Typology of Existential Vulnerabilities and Exposures for Existential Risk Research." Futures. https://www.researchgate.net/publication/324688255_Governing_Boring_Apocalypses_A_New_Typology_of_

References

Existential_Vulnerabilities_and_Exposures_for_Existential_Risk_Research

Mallah, R. 2017. "The Landscape of AI Safety and Beneficence Research: Input for Brainstorming at Beneficial AI 2017." Future of Life institute.
https://futureoflife.org/landscape/ResearchLandscapeExtended.pdf

McLarty. Thomas F; Ridge, Thomas L. 2014. Securing the U.S. Electrical Grid. Center for the Study of the Presidency & Congress.
https://www.thepresidency.org/sites/default/files/Final%20Grid%20Report_0.pdf

McReynolds, J. 2017. "China's Evolving Military Strategy." Jamestown Foundation.
https://jamestown.org/product/chinas-evolving-military-strategy-edited-joe-mcreynolds/

MIRI.2018. "Aligning advanced AI with human values." MIRI.
https://intelligence.org/research/

Nitzberg, Mark, Groth, Olaf.
Solomon's Code: Humanity in a World of Thinking Machines. 2018.

Onynshkevych, B. 2018. "Active Interpretation of Disparate Alternatives," DARPA
https://www.darpa.mil/program/active-interpretation-of-disparate-alternatives

OpenAI, 2018. "OpenAI Charter". OpenAI.
https://blog.openai.com/openai-charter/

Partnership on AI. 2018. "Tenets". Partnership on AI.
https://www.partnershiponai.org/

Peterson, C., Miller, M., Duettmann, A. 2017."Cyber, Nano, and AGI Risks: Decentralized Approaches to Reducing Risks." UCLA Risk Colloquium, 2017.
https://ai.google/research/pubs/pub46290

Potember, 2017. "Perspectives in Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD." DoD.
https://fas.org/irp/agency/dod/jason/ai-dod.pdf

Rashid, Fahmida Y. 2015. "Cyber Attack on Power Grid Could Top $1 Trillion in Damage: Report." Security Week, July 16, 2015.
http://www.securityweek.com/cyber-attack-power-grid-could-top-1-trillion-damage-report
Russell, S. Dewey, D., Tegmark, M. 2018. "Research Priorities for Robust and Beneficial AI." Future of Life Institute.
https://futureoflife.org/data/documents/research_priorities.pdf

References

Seffers, G. 2018. "AFRL Anticipates Arrival of Neuromorphic Supercomputer." AFCEA.
https://www.afcea.org/content/afrl-anticipates-arrival-neuromorphic-supercomputer

Shuhiyu, D. 2018. "China forms 1st AI Alliance." ChinaDaily.
http://www.chinadaily.com.cn/business/2017-06/21/content_29833433.htm

De Spiegeleire, S., Maas, M., Sweijs, T. 2017. "AI And The Future of Defense." The Hague Center for Strategic Studies.
https://hcss.nl/sites/default/files/files/reports/Artificial%20Intelligence%20and%20the%20Future%20of%20Defense.pdf

Tuskey, K. "The Biggest Crypto-Partnership The World Has And Will Ever See." Medium.
https://medium.com/@keithtuskey/the-biggest-crypto-partnership-the-world-has-and-will-ever-see-2903f65ab107

Wikileaks, 2013. "Nations Buying As Hackers Sell Computer Flaws."
https://wikileaks.org/hackingteam/emails/emailid/96008

Yampolskiy, R. 2015. "Artificial Superintelligence: A Futurist Approach." CRC Press.
https://www.amazon.com/Artificial-Superintelligence-Futuristic-Roman-Yampolskiy/dp/1482234432

Yampolskiy, Roman. 2018. Beyond MAD?: The Race for Artificial General Intelligence.
https://www.itu.int/en/journal/001/Documents/itu2018-9.pdf

Ye, J. 2017. "China tightens Great Firewall by declaring unauthorized VPN servers illegal." South China Morning Post.
https://www.scmp.com/news/china/policies-politics/article/2064587/chinas-move-clean-vpns-and-strengthen-great-firewall

FORESIGHT INSTITUTE

Artificial General Intelligence: Coordination & Great Powers

A white paper based on the
2018 Foresight Institute Strategy Meeting on AGI