FORESIGHT INSTITUTE

Artificial General Intelligence: Toward Cooperation

Allison Duettmann, Foresight Institute - corresponding author
Alan Karp, Earth Computing
Anthony Aguirre, Future of Life Institute
Baeo Maltinsky, Median
Brian Tse, Future of Humanity Institute
Christine Peterson, Foresight Institute
Colleen McKenzie, Median
Gaia Dempsey, 7th Future
Jeffrey Ladish
Jim O'Neill, SENS Research Foundation
Jingying Yang, Partnership on AI
Lou de Kerhuelvez, Foresight Institute
Mark Miller, Agoric
Peter Eckersley, Partnership on AI
Pindar Wong, Hong Kong Smart Contracts Initiative
Robin Hanson, George Mason University
Sarah Constantin, Longevity Research Institute
Rosie Campbell, Partnership on AI
Tom Kalil, Schmidt Futures
Tasha McCauley, OpenAI

Table of Contents

Participants

Aguirre, Anthony	Future of Life Institute
Anderljung, Markus	GovAI, Future of Humanity Institute
Bach, Joscha	AI Foundation
Belfield, Haydn	Center for the Study of Existential Risk
Bowerman, Niel	80,000 Hours
Byrd, Chris	Tsinghua University
Braley, Ryan	Lightbend
Bourgon, Malo	Machine Intelligence Research Institute
Campbell, Rosie	Partnership on AI
Cihon, Peter	Future of Humanity Institute
Constantin, Sarah	Longevity Research Institute
Constantinescu, Mirona	
Critch, Andrew	Center for Human Compatible AI
Cuperman, Miron	Base Zero
Cussins, Jessica	Center for Long-Term Cybersecurity
De Kai	Hong Kong University of Science & Technology
de Kerhuelvez, Lou	Foresight Institute
Dempsey, Gaia	7th Future
Ding, Jeffrey	Future of Humanity Institute
Duettmann, Allison	Foresight Institute
Eckersley, Peter	Partnership on AI
Flidr, Ales	Deepmind
Garrick, John	John Garrick Institute for the Risk Sciences
Girshovich, Dan	Rigetti Computing
Hadshar, Rose	Future of Humanity Institute
Hanson, Robin	George Mason University
Hay, Nick	Vicarious
Karp, Alan	Earth Computing
Kalil, Tom	Schmidt Futures
Katz, Eddan	Center for the Fourth Industrial Revolution, WEF
Kasewa, Tanya Singheef	Future of Humanity Institute
Ladish, Jeffrey	
Leike, Jan	DeepMind
Leung, Jade	Future of Humanity Institute
Mallah, Richard	Future of Life Institute
Maltinsky, Baeo	Median
Maini, Vishal	DeepMind
McCauley, Tasha	OpenAI
McIntyre, Peter	80,000 Hours
McKenzie, Colleen	Median
Miller, Mark	Foresight Institute, Agoric
Nixon, Jeremy	Google Brain
O'Neill, Jim	SENS Research Foundation
Peterson, Christine	Foresight Institute
Rakova, Bogdana(Bobi)	Accenture Responsible AI
Ross, Nicole	Centre for Effective Altruism
Taylor, Jessica	Median
Tse, Brian	Center for the Governance of AI, Oxford
Vassar, Michael	
Vance, Alyssa	Apprente
Ward, Liz	John Garrick Institute for the Risk Sciences
Wong, Pindar	Hong Kong Smart Contracts Initiative
Yang, Jingying	Partnership on AI

Executive Summary

Allison Duettmann
Foresight Institute
a@foresight.org

This report summarizes the main findings of the 2019 AGI Strategy Meeting on "Toward Cooperation: Framing & Solving Adversarial AGI Topics," held in San Francisco on June 20, 2019. This annual meeting is hosted by Foresight Institute, a 501(c)3 non-profit organization dedicated to selectively advancing technologies for the long-term benefit of life.

The 2017 meeting in this series (report) focused on drafting policy scenarios for different AI time frames, and was followed by the 2018 meeting (report) that focused on increasing coordination among AGI-relevant actors, especially the US and China. The 2019 meeting expanded on this topic by mapping concrete strategies toward cooperation. This includes both reframing adversarial coordination topics in cooperative terms and sketching concrete positive solutions to coordination issues. The meeting gathered representatives of major AI and AI safety organizations with policy strategists and other relevant actors with the goal of fostering cooperation amongst global AGI-relevant actors but also more directly amongst participants to contribute to a flourishing long-term community.

A group of participants presented their recent efforts toward a more cooperative AI landscape, followed by discussion in small groups. While discussions followed the Chatham House Rule, a high-level summary of the sessions is available in this report. Please find the sessions, grouped according to topics, with their respective action items below, followed by longer session summaries. We welcome questions, comments, and feedback to this report.

Executive Summary

TOPIC	NEXT STEPS
1. THE NEED FOR COOPERATION IN AI GEOPOLITICS	
Improving Coordination with China to Reduce AI Risk	Develop ideas for common knowledge and coordination mechanisms for reducing potential racing dynamics/ safety-performance trade-off
Chinese Belt & Road Blockchain as Safety Mesh for AI	Silicon shock US China: As the system of dependency on silicon is pivoting, we need to explore new paradigms
AI Impacts on Escalation Paths to Nuclear Conflicts	Build partnerships between industry, government, and academic circles to create common knowledge about AI-nuclear escalation risks
2. THEORETICAL CONSIDERATIONS AFFECTING AI COOPERATION	
Security: The Mess We're In, How We Got Here, and What to do About It	When building a system that involves permission management, enforcing least privilege at fine granularity must be part of the architecture from the start
Implications of AI Timelines Study for AI Coordination	Explore scenario planning, no regrets scenarios, and new hardware paradigms
Strategic Considerations for Openness in AI Research	PAI to possibly look into setting up a 'red team' service for unintended consequences of research and explore "unintended consequences" section added as a norm for publications
3. SHORT TERM AI NEEDS AND APPLICATIONS AS GROUND FOR COOPERATION	
Building Out National AI Infrastructure	Identify areas where investments in AI infrastructure would have a large impact, such as training data
Using ML to Solve Concrete Societal Problems	Research and quantify trends in performance improvements in surveil-and-optimize AI systems for increasing output (e.g. agriculture, manufacturing)
Status Report on Projects from BAGI 2019	Multiple cooperation-focused projects from BAGI need additional stewarding. Contact FLI to collaborate on or lead specific projects
4. ONGOING PROJECTS SEEKING TO INCREASE COOPERATION	
AI Safety Communication: Stakeholders, Narratives, Projects	Help PAI with developing a cross cultural framework; help Foresight with AI chapter of the Existential Hope primer
A Framework for AI Fiduciary Agents	Study principal-agent literature. Define criteria for fiduciary agent application. Create prototypes at hackathons
5. REFRAMING AI TRAJECTORIES IN COOPERATIVE TERMS	
Intelligent Voluntary Cooperation	Contact Foresight Institute for more information on the book
Policy for Decentralized Take-Off Scenarios	Explore robots-take-most-jobs insurance

1. The Need for Cooperation in AI Geopolitics

Improving Coordination with China to Reduce AI Risk

Session chaired and summarized by Brian Tse

Overview

One of the key recommendations from Foresight's Artificial General Intelligence: Coordination & Great Powers report last year was to "encourage early cooperation on concrete issues with lower stakes to create precedent and infrastructure for later cooperation." In 2019, there are significant efforts and a willingness to push for AI governance, safety, and international cooperation from the Chinese government and industry. This represents a timely opportunity to discuss concrete issue areas in which coordination between Chinese and foreign actors, especially between the leading AI developers, can be explored.

A forthcoming report from Brian Tse at the University of Oxford's Center for the Governance of AI discusses promising issue areas and approaches to coordination involving Chinese AI developers. It recommended that the coordination effort should focus on four issue areas, namely: reducing risks from accidents, misuse and racing dynamics, and committing to broadly distributed benefits. In terms of the approaches to coordination, there are three promising areas, including lab-to-lab partnerships; industry and academic consortia; and expert communities.

Benefits & Advantages

Such an effort would serve as a direct answer to the calls from Chinese political and business elites as well as major Western institutions. In a congratulatory letter delivered to the 2018 World AI Conference in Shanghai, Chinese President Xi Jinping stated that "China is willing to join hands with other countries to promote the development of artificial intelligence, ensure security, and share fruitful results." Similarly, Baidu's CEO Robin Li has said that what the two countries need is to explore cooperation with mutual benefits while continuing with healthy competition between the companies. Baidu

became the first Chinese member at the Partnership on AI. Demis Hassabis, co-founder of DeepMind, has claimed that "the coordination problem is one thing [we should focus on now]. We want to avoid this harmful race to the finish where corner-cutting starts happening and safety gets cut ... That's going to be a big issue on a global scale."

The focus on AI developers, who are actors that attempt to develop machine learning (ML)/AI systems, allows the coordination to be facilitated by interactions between firms, academia, and the wider civil society. This can potentially continue despite the ongoing tensions between nation-states, especially because the AI research community is bounded by shared best practices and values across the world. The focus on technical exchange and coordination mechanism in reducing AI risks is also mutually beneficial and much less ideologically or politically controversial.

Obstacles & Risks

At a time of intensifying rivalry between China and the United States, a recent Brookings' report on AI summarizes the American policies that have arisen from some of these concerns (as well as the broader emergence of great power competition): "tightening screening of foreign investments in core technologies, scrutinising Chinese academic exchanges, applying targeted tariffs to reduce China's competitiveness in key sectors, increasing prosecutions of Chinese actors involved in economic espionage, and investing greater resources in counter-intelligence operations." It exemplifies a fundamental worry that even well-intended, technical exchange might be insufficient to avoid unforeseen applications or outcomes that undermine the national security of other countries.

Next Steps

Outside of China, there are more than 10 organizations contributing to the beneficial and positive development of AGI. This concern is increasingly understood by Chinese institutions. However, there are relatively few organizations and full-time teams of researchers working on this problem within the country. Moreover, there is no existing mechanism to coordinate a mutually beneficial risk reduction effort between Chinese and international actors. Therefore, setting up an initiative to improve Chinese-foreign partnership in reducing AI risks seems to be an important and unexplored effort. This was proposed by Brian Tse, the chair of the session.

Further Investigation

There are several categories of research questions that come out from the discussion. The first one relates to the institutions and mechanisms for such coordination to take place, such as the most effective international forum to support China-US cooperation on AI. The second category is concerned with capabilities, such as the likelihood for China to develop a megaproject on AGI. The third category relates to ethics. For example, what is the expected trajectory of the safety research field in China, and would these safety techniques be shared globally? What organisations in China are the most forward-looking in their thinking about ethical aspects of AI?

Chinese Belt & Road Blockchain as Safety Mesh for AI

A session chaired by Pindar Wong, Hong Kong Smart Contracts Initiative, and summarized by Lou de Kerhuelvez, Foresight Institute

Overview

International trade traditionally benefits global economies, with the economies of China and the US serving key leadership roles in the mutual supply and demand of tangible goods, intangible services and intellectual property. President Xi Jinping's 'Belt and Road Initiative' is in China's Constitution, with its data-driven economy now spending more on importing semiconductors than on importing crude oil (China Semiconductor Industry Association notes USD 260 Billion vs USD 162 billion in 2017 respectively). However, current trade tensions are fostering negative narratives such as 'decoupling' and alternative projects (NB: In October 2019, post-event, the US announced its Blue Dot Network and President Xi publicly emphasized the importance of Blockchain Technology research to China's national development).

In the perspective of fostering positive-sum game scenarios, this session presents some design principles and functional capabilities of a blockchain-based system to transfer titles, record permissions and provenance logs; so as to automatically track and trace the safe flow of goods and services, between businesses, across national borders: an AGI-enhanced border security and digital dispute resolution system. This system adopts a two-layered architecture -- separating the common commercial need of legal signatory accountability for cryptographic key management, from the different functional reporting requirements found across a myriad of commercial industries, cultures and customs contexts.

In the end, the idea is to increase the transparency of regulatory customs-tariff compliance and collection, to reduce trade friction, by lowering those common costs related to establishing the legal 'facts' used in resolving cross-border trade disputes. The promise of deploying AGI here is to automatically detect and flag trade anomalies such that legal accountability can be traced across the length and breadth of increasingly complex manufacturing and trade networks.

The opportunity, and yearning, is for both China and the US to lead the development of machine learning quality control, to move from ad-hoc heuristic exploration to a better-understood science, through considering the introduction of some form of 'Proof of Training': namely, the systematic and deliberate development of machine learning process provenance in the course of developing machine learning algorithms. By working together to establish open standards for the mathematical evidentiary treatment of data (a side effect of which is the generation of an agreed Proof-of-Training (PoT) token), a 'best-common-practice' for AGI procedural hygiene can be developed for the training, development, and testing of relevant datasets used in proprietary machine learning algorithms.

If it takes a proverbial village to raise a child, it might take a planet to raise an AGI.

Benefits and Advantages

The main benefit here is mutually assured cooperation, and the shift from deploying an accidental internet of threats to deliberately incentivize an internet of trust. Indeed, this system is aimed at mediating trust, increasing selective transparency, detecting trade fraud and expediting digitized customs declaration, with benefits in traditional 'CIQ' (Customs Inspection and Quarantine) processes and commitments. An avenue to avoid AI races scenarios and foster US-China collaboration is the development and implementation of 'Proof-of-Training' algorithms. This will be especially true as the next generation of AI-optimized chips emerge both in China and the US (e.g. Google's TPU, Bitmain's Sophon series, Graphcore (UK)). AGI can be applied in automated online trade dispute resolution:

c.f. Alibaba's application of its datasets and algorithms to help drive Hangzhou Internet Court for China's domestic e-commerce ecosystem. An opportunity exists for collaboration to adopt a scientific approach of Proof-of-Training in the field of international customs inspection and quarantine (CIQ) and digitized custom processes between China/US, and elsewhere, in both the trade in tangible and intangible goods and services.

Risks and Obstacles

The question of legal accountability: where does one draw the checks and balances?
Blockchains are inherently 'slow' networks and don't scale very well. Incentives to participate needs to be created, together with suitable punishments for consensus deviation.

Next Steps

Encouraging the next-generation of AI-optimized Chips to include mathematical evidentiary computation and process hygiene to be rewarded with 'Proof-of-Training' (PoT) tokens to provide a 2nd order signaling to detect runaway/stumbling on AGI.

Encouraging discussion on the scope, costs, and kinds of algorithmic transparency of a PoT token to power cross-border M2M(Machine-to-Machine) economies in trust-minimized networks.

Further Investigation

What parity, or gap, in PoT token production, would be considered tolerable is a matter of mathematical research and practical engineering finesse (c.f. Traditional measures of advancement in supercomputing). The purpose of encouraging PoT algorithmic research focusing on Machine Learning process hygiene design is to deliberately and collectively benchmark and incentivize the detection of run-away/breakthrough AGI development. Any sudden and unexpected spike in PoT token production would indicate an emerging AGI-gap, in spite of compliance with 'established' training hygiene norms, such that non-technical steps can be triggered to quickly de-escalate, pause or narrow any AGI-gaps that emerge.

AI Impacts on Escalation Paths to Nuclear Conflicts

Session chaired and summarized by Jeffrey Ladish

Overview

Advances in military-relevant AI technologies threaten to worsen risks of accidental nuclear escalation. Perhaps the greatest risk comes from the uncertain effectiveness of these capabilities, both in theory and in application. Despite their power, these technologies should not be seen as a "game changer," but should be understood in the context of existing and in-development weapons and deployment capabilities. While we believe the most likely effect of these technologies will be to increase the risk of inadvertent escalation, there is an opportunity to use these technologies to accomplish the opposite, and reduce these risks.

Since the invention of nuclear weapons, competition for nuclear advantage has driven advances in three areas. The first is the ability to reliably, accurately, and quickly deliver nuclear weapons to their targets. The second is the ability to effectively monitor nuclear developments in other countries, including the development and tests of nuclear weapons, the deployment of nuclear weapons, and, crucially, the launch or use of nuclear weapons. The third category, the ability to directly defend one's

weapons, and, crucially, the launch or use of nuclear weapons. The third category, the ability to directly defend one's country from nuclear attack, has also advanced, though carries far less relevance to a great power conflict due to the fundamental difficulty of missile defense. AI technology will impact all three areas.

The greatest risks to strategic stability come from uncertainties about all three categories of development, with the greatest risks stemming from the first, the ability to reliably, accurately, and quickly deliver nuclear weapons to their targets. On the other hand, concrete advances in the second category, the ability to effectively monitor nuclear developments of states around the world and detect weapons tests and missile launches, has the potential to improve strategic stability. Improving the reliability of launch detection systems would reduce the risk of accidental escalation, and improving the detection of weapons development could improve confidence in nuclear arms control agreements.

Benefits & Advantages

The greatest potential advantage for AI technology lies in improving detection capabilities -- that is, the ability to effectively monitor nuclear developments in other countries, including the development and tests of nuclear weapons, the deployment of nuclear weapons, and, crucially, the launch or use of nuclear weapons. More reliable missile launch detection capabilities could reduce the risk of accidents caused by the current launch detection systems which have previously exhibited near-disastrous false positives. Greater monitoring capability of nonnuclear nations can help prevent further nuclear proliferation and lead to better enforcement of the Treaty on the Non-Proliferation of nuclear weapons (NPT). Improved ability to monitor the nuclear developments of current nuclear states may allow for greater confidence in existing and future arms control agreements.

Risks & Obstacles

The greatest risks to strategic stability and the prevention of nuclear escalation come from two main factors. The first is uncertainty of nuclear defensive and offensive capabilities, which can worsen arms races and increase pressure to act in the fog of war. The resulting missile build up in the US due to the perceived "missile gap" during the Cold War demonstrates how inadequate intelligence about another state's nuclear capabilities can worsen arms race dynamics. If deployment of existing AI technologies lead to improved intelligence, reconnaissance, and surveillance (ISR), the net effect should be to reduce the states' uncertainties regarding each other's nuclear capabilities.

The second factor is the ability for AI technologies to increase the speed and stealth of weapons by enabling enhanced autonomous capabilities. We expect that advances in autonomous vehicles, especially drones, will lead to systems with greater reaction times and an increased capacity for rapid conflict escalation. This would decrease the time for military leadership to react in a crisis and increase the incentive for preemptive escalation. Furthermore, a natural response to the threat of enemy autonomous weapons systems will be an arms race in these types of systems.

Next Steps

The application of military-relevant AI technologies is full of uncertainty. We cannot foresee exactly how the technologies will develop nor understand their full impact. The same military uncertainties that heighten risk of accidental escalation also make the development of public policy on these topics quite challenging. We believe the only way to make progress in this domain is to improve the quality of research and discussion and the pipeline that connects this discussion to actual policy.

When nuclear weapon systems were first developed in the 1940s and 1950s in the United States, public and private collaboration was much stronger than it is today. Unlike that era, most advances in AI technologies now come from private companies, often in collaboration with academic institutions. As a result, policy makers in the US government and military suffer from a lack of expertise in the technology their policies target. We recommend that researchers and private institutions with active safety research programs and a strong commitment to the safe application of AI technologies establish relationships with policy makers to improve mutual understanding of technology and policy challenges.

Ultimately, reducing risk from nuclear escalation has a large impact on all countries in the world. While all militaries have incentives to hide and protect strategic secrets, there are also reasons to collaborate on understanding the strategic ramifications of new technologies. No country in the world would benefit from nuclear escalation. We hope that states will choose to develop AI technologies in areas where they will improve strategic stability while working towards international agreements to limit destabilizing applications of these technologies.

Further Investigation

We recommend a thorough study of existing and near-horizon autonomous weapons systems and their impact on nuclear escalation paths. So far, this section has avoided discussion of AGI due to the speculative nature of AGI systems in contrast to the comparatively better specified nature of current AI technologies and their application to nuclear defense and offensive systems. The one area in which AGI appears relevant is how perceptions of AGI development may affect nuclear stability and the deterrence of nuclear first strikes. Even if AGI capabilities are far out, the perception of a state's nearness to decisive strategic advantage via AGI capabilities could worsen strategic stability. We recommend cautious investigation of the impact of perceived AGI capabilities on strategic stability.

Security: The Mess We're In, How We Got Here, and What to do About It

Session chaired by Alan Karp

This session was a talk by Alan Karp designed to get AI and AI safety researchers up to speed on the precarious state of our global computational infrastructure and the criticality of improved computer security. The slides of the presentations are available for download here, with more detailed introductions to Object Capabilities available here and here. Relevant talks include Alan Karp on Writing Applications that are Easier to Defend than Attack, about The Virus Safe Computing Initiative, and Mark S. Miller on Computer Security as the Future of Law.

2. Crucial Theoretical Considerations Affecting AI Cooperation

Implications of AI Timelines Study for AI Coordination

Session chaired by Baeo Maltinsky and Colleen McKenzie, summarized by Colleen McKenzie

Overview

The coordination of multiple agents in the AI space, and analysis of that coordination, both benefit from an understanding of the timing of events. An uneven distribution of technological progress makes forecasting difficult, but continued refinement of predictions and expectations is important.

Benefits & Advantages

The potential outsize impact of sophisticated AI obligates a thoughtful allocation of resources toward ensuring that the impact is positive. Where these resources are most effectively allocated will depend on what progress is expected, when, and in which areas. Risks of near-term developments may pose risks that are best mitigated by non-technological means—for example, the political risks associated with synthesized audiovisual "deep fakes"—but must be weighed against mitigating risks from more significant future advances. Near-term forecasting may also provide opportunities for better calibration of forecasting abilities: predictions of imminent breakthroughs may be tempered by slow progress, and skeptics may update in the face of unexpected results.

Further, because AI is not the only global catastrophic risk we face, allocation of resources itself is a coordination problem. Expectations of when major developments will occur will be important parameters for attempts to solve this problem.

Obstacles & Risks

Attempts at accurate forecasting constitute a study in working around and despite ubiquitous obstacles. The last 70 years of progress in AI has been accompanied by constant predictions of the pace of advances, most of which have been incorrect. Progress in the space is lumpy: periods of significant discovery and application are interspersed unpredictably with AI winters, and important theoretical developments sometimes go unrecognized and unrealized for years. Predictions for AGI are particularly difficult as they are confounded by disagreements on the definition of AGI itself.

Prediction efforts that use historical data to predict future advances can take this history of forecasting into account as an additional parameter, but the difficulty of weighting the importance of past discoveries makes this difficult: it's possible that a future major advance in AI capabilities will rely on past advances that hadn't yet found their most effective application and thus are misjudged as relatively insignificant.

Forecasting in other domains is similarly difficult: for example, we have no evidence that individuals correctly predicted the development of the atom bomb even just five nears before its achievement.

Next Steps

Scenario planning can be an effective tool for forecasting, particularly for near-term risks. Enumeration and comparison of potential impacts of progress in AI could provide concrete enough predictions to allow weighting and tradeoffs in the face of considerable uncertainty about outcomes. Job loss from increased automation capabilities, for example, has been agreed upon as likely to some given degree, but forecasts of the degree diverge considerably. Models of the concrete effects of specific technological developments and their interactions would allow for inspection of how forecasters disagree on the relevant mechanisms of action.

AI forecasting will benefit from more and better forecasting in adjacent domains. Major advances in hardware used for computing, in particular, could cause step-function changes in machine intelligence capabilities without direct algorithmic improvements. Optical computing is the most promising advance in the space evident at the present time.

Further Investigation

Given the difficulty of forecasting with certainty, it seems worthwhile to find policies applicable to the broadest possible set of outcomes. Such "no-regrets policies" could be determined based on a collection and probability distribution of possible outcomes even if claims about the distribution were weak.

Strategic Considerations for Responsible Publication Norms in AI Research

Session chaired by Peter Eckersley, summarized by Rosie Campbell

Overview

In early 2019, OpenAI made headlines with its advanced language generation model GPT-2. What attracted attention however was not just the technology itself, but also OpenAI's decision to withhold from publishing the full model due to potential malicious applications, such as spam or fake news. While some praised the cautious approach, and the discussions around

responsible publications norms that it spawned, others expressed concerns about the downsides of closed research practices.

The salient issue is not whether OpenAI made the right call in this particular instance, but the fact that as Machine Learning (ML) systems become increasingly advanced, researchers are frequently going to face these kinds of difficult decisions about what to do with research that carries significant risks.

We are already seeing examples of harmful consequences of ML technology. Engagement optimization algorithms on social media are often accused of causing political polarization and altering our attention spans, and we're starting to see a variety of nefarious uses of convincing 'deepfake' videos. There is increasing urgency for the ML community to establish the right norms and precedents to maximize the benefits of openness while mitigating risks as the capabilities of these systems increase. Researchers and organizations have prudently begun to request guidance on responsible publication, openness, and precautionary review processes.

Benefits & Advantages

Norms and processes that successfully balance openness with risk reduction would allow us all to safely enjoy the numerous benefits of advanced ML systems, while avoiding protectionism. It would also help give researchers confidence in their work and alleviate anxiety about possible harms.

Changing community norms can be a difficult process, but there is reason for optimism: The ML community already has a history of successfully changing publication norms when it began publishing in open access journals like arxiv, going against significant pressure to publish in traditional closed academic journals.

It's also important to recognize that the choice is not simply 'publish or don't publish'. Openness is a spectrum, and this affords us many possible options for maximizing the benefits while mitigating risks. Examples can include delaying publication (so that research can be done on harm mitigation measures before release), keeping parts of the research private but releasing the rest, and limiting the release to certain groups. Similarly, there are many possible options for precautionary review processes, from lightweight checklists to internal or external review boards, and more. This diversity of options available means that it is possible to create bespoke release plans suited to the particular circumstances of each research advance.

Risks & Obstacles

Striking the balance between keeping research as open as possible while keeping it closed enough to minimize risks of misuse is a delicate process. Too open, and we risk information hazards, accidents, unintended consequences, and malicious use. Too closed, and we risk losing the incredible benefits of open science and peer review, duplicating effort, creating barriers to entry for aspiring researchers, and intellectual bubbles. There may also be second-order effects, such as stalling the development of harm mitigation measures (e.g. research to detect deepfakes), or organizations using the guise of social responsibility to keep their work secret and protect IP.

One obstacle is whether it is even possible for researchers to accurately anticipate the harm their research could cause. It's unlikely that researchers of multi-armed bandit algorithms in the late 20th century could have predicted that decades later these algorithms would one day power social media, affecting our cognitive attention and arguably destabilizing democratic societies. But perhaps if researchers had been encouraged to think creatively about the risks of their work they may have anticipated the application of bandit algorithms to advertisements, and from there anticipated their ability for mass manipulation. Recognizing this earlier could have given us much more time to prepare for today's technology.

Encouraging researchers to anticipate and communicate the potential harms of their work also comes with its own set of risks. Being more upfront about potential misuse could be an information hazard and give malicious actors dangerous ideas about how to use the technology to cause harm. There are also unanswered questions around whose responsibility

it is to anticipate the risks: technical researchers may be the expert in their own work, but they may not have the social science expertise needed to anticipate its broader impact on society.

We should also be mindful of the consequences of introducing new norms or processes that can introduce incentives that interact unpredictably with existing structures. Overly cumbersome precautionary review processes could frustrate researchers and cause them to try and find ways around it. Asking them to add a discussion of risks to their grant application and papers could be seen as just additional bureaucratic overhead and be reduced to a box-ticking exercise.

Next Steps

The primary next step is to liaise with members of the ML community on the issues of responsible publication, openness, and precautionary review. This is currently underway and is being coordinated by the Partnership on AI with the input of a diverse array of partner organizations. As well as soliciting opinions and expertise, this step also involves looking at norms in other fields such as cyber-security and biotechnology which have had to face similar concerns and have developed risk mitigation strategies, some of which may also benefit the ML community.

Further Investigation

As discussed, there are a variety of ways to approach responsible publication, openness, and precautionary review. Some additional ideas that seem particularly promising include:

- Encouraging researchers to include an 'unintended consequences' or 'risks' section when applying for funding or submitting to journals and conferences. This could either be directly enforced (e.g., grant reviewers and journals won't consider work that doesn't include this) or indirectly (social approval or disapproval from research peers).
- Creating a 'red team' service to help researchers, peer reviewers, and grant reviewers assess the potential harms of a research project or technology. This could be a diverse panel of experts skilled in anticipating unintended consequences and second-order or flow-through effects of technology on society. The panel could include experts in the technical research area, experts from social sciences, and even science fiction writers.

More ideas like this will likely come out of the liaison process with the ML community. Of course, the ideas themselves may have unintended consequences, so further investigation will be required to ensure they are implemented effectively and don't create perverse incentives or have otherwise negative effects.

3. Short-term AI Needs and Applications as Ground for Cooperation

Building Out AI Infrastructure

Session chaired by Tom Kalil, summarized by Jingying Yang

Overview

Any solutions for cooperation on adversarial AGI topics or building AGI will be built on AI infrastructure, which encompasses hardware (GPUs, TPUs), labeled data sets, AI test beds (e.g. for AI and cyber physical systems), research infrastructure (between AI and science), environments (for RL), benchmarks, programming languages and libraries, pipelines to connect datasets from different local jurisdictions, resources for people who will build any of the above infrastructure, and beyond.

In many cases, AI infrastructure is a public good, which runs into the same tragedy of the commons dynamics of many public goods. There is a large imbalance between the high number of people who could use or benefit from the infrastructure compared to the low number of people willing to do the hard work of building it. As a collective, we are underinvesting in AI infrastructure relative to its importance. With so many diverse types of AI infrastructure, which all have interdependencies, one approach to accelerate progress is to unbundle two processes:

- Identifying areas where investment in AI infrastructure would have high ROI; and
- The hard work of building the AI infrastructure.

It would be worthwhile to first undertake a systematic agenda-setting process to research the costs, benefits, and synergies from different types of AI infrastructure investments, which would help foundations, government agencies, and nonprofits to direct their funding efforts to the most promising projects and thus accelerate progress in the highest priority areas. This process should also study existing resources to see whether they can be easily repurposed as AI infrastructure with only relatively small amounts of effort to unlock enormous benefits. For example, the US Department of Defense owns over a million pathology slides, which could become the biggest dataset for radiology ML models if digitized. There are many more hidden treasures like this waiting for discovery.

Building datasets is both high impact and relatively neglected. A recent study showed that many advances occur 18 years after the publication of the inaugural paper, but only 3 years after a relevant dataset became available, which shows that data is the rate limiting step for technological advancement. However, most organizations underinvest in building data resources, which makes it a more promising opportunity for active investment.

Next Steps

To build out a holistic investment plan for the field, there are some key questions to answer:

- How do we identify people who could generate plausible hypotheses about what types of investments in infrastructure we should be supporting?
- What is the process we should use for identifying the types of investments we should be making?
- Are there elements of AI infrastructure missing from the above list?
- Are there specific examples of investments we should be making within each infrastructure type?

Once types of AI infrastructure are prioritized, projects need funding, talent, and time. Having a public list of prioritized AI infrastructure projects would also lower barriers for attracting all three.

Using Machine Learning to Solve Concrete Societal Problems

Session chaired by Sarah Constantin, summarized by Allison Duettmann

Overview

This section summarizes promising near-term applications of AI to fundamental human needs, like food, water, shelter, safety from violence, and safety from disease. Solving those large scale societal needs often requires more data than one company administers, providing a possible context for increased cooperation.

Benefits & Advantages

We can expect further near-term progress from neural nets on highly multi-variable data sets with non-linear relationships, including images, video, text, audio, user behavior, and logs of sensors. The bottleneck to near-term positive impact from AI will be partly determined by how many things we can we make into a problem of cheap sensors and analysis. In theory, anything that can be monitored via cameras or satellites lends itself to optimization. A few examples include:

- Manufacturing is still much less roboticized than commonly assumed in the press. Increased monitoring of the assembly line process is needed to determine which parts can be increasingly roboticized.
- The biomedical world could be improved by a factor of 10 by making any problem that is currently solved as a genetics problem, e.g. DNA sequencing and RNA sequencing, an imaging problem. In principle, anything that a cell biologist can see under a microscope could be processed by an imaging program.
- Shipping and logistics ought to be amenable to the same kind of game design as video games.
- In agriculture, monitoring is key for productivity. Optimizing particular batches of fermenting microbes that produce a chemical reaction requires monitoring, just like locating large scale overfishing. Individual plants can

be monitored at very high granularity to improve performance and whole areas of crop lands can be monitored to determine where to improve the water supply. Up to 20% improvement rates are feasible in crop fields from no new technology except for fine-grained optimization.

- Imaging-based optimization has potentially positive impacts in geoengineering for climate change. For instance, iron fertilization of water to change the local climate is a problem that requires a large fleet of sensors.

Risks & Obstacles

We need to answer the following questions in order to better understand the risks and obstacles arising from machine learning applications for near-term problems:

- Is fine-grained optimization the right application of AI? Finding anomalies and potential dangers is a much harder problem than using AI to make an initial pass and relying on humans to handle the enormous variety of edge cases.
- Are our algorithms actually improvements over the status quo? Surprisingly few papers introduce a specific new algorithm that is significantly better than the genetic algorithms of the 80s, e.g. random forest, just with more processing power. Comparisons across algorithms are hard because we lack good universal benchmarks for comparison.

Next Steps

Even in cases in which machine learning ought to accelerate progress in principle, progress at scale is impeded by political problems like data integration. Collaboration across departments, teams, and organizations does not evolve organically within the kinds of organizations that do most of the agriculture and aerospace engineering. AI specialists could aid support data integration in those areas.

In addition, we need better market analysis to determine how much output improvement is possible via AI optimization. While benchmarks exist for classifying accuracy, we lack benchmarks for transfer learning in various applications because every organization guards their own data lakes. A welcome innovation would involve introducing benchmarks that are based on an output metric that is open to different approaches that can be compared. For instance, answering questions like “how much crop yield improvement is possible, given specific satellite data points?” would give us a good understanding of what the state of the art can and cannot do.

In sum, we need to move from individual companies solving specialized problems to the ability to compare capability across the board. This first step in allowing us to reap near-term benefits from AI requires considerable cross-company analysis and could be explored as an incentive to increase cooperation.

4. Ongoing Projects Seeking to Increase Cooperation

Status Report on Projects from BAGI 2019

Session chaired and summarized by Anthony Aguirre

Overview

The 2019 Puerto Rico meeting on Beneficial AGI resulted in over two dozen concepts for projects of various scales conceived as groundwork to put in place in the near future to increase the probability of an overall positive outcome at the advent of general/transformative AI. Some of these projects were already in progress as of the B-AGI meeting, some have been picked in the meantime by various groups, and some are still at concept stage. From the list, the currently most highly-developed projects are:

a. Support for an international agreement controlling and curtailing the use and development of lethal autonomous weapons (led by FLI);

b. A collaborative network and platform for probabilistic predictions of AI and other technological progress (led by FLI, Metaculus, Parallel); and

c. A legal and social instrument to allow AGI research companies to credibly commit to sharing their profits above some (very high) level (led by GovAI).

Each of these is mature but is in need of further support and coordination. In-development projects that have begun but still require additional personnel, funding, and intellectual input include:

d. Well-developed ideas and seed institutions for using the profits from the "windfall clause" (project (d) above). In the event of its triggering, the clause would create by far the largest non-profit/charitable institution in history, so concomitant care must be taken in that institutions's structuring;

e. A framework for standards, or perhaps legal framework, as to how an AI agent can have a fiduciary status, i.e.

take a person's (or organization's) best interest as their own best interest, in a clear and transparent way; and

f. A proposal to create major government-funded CERN-like center for advanced collaborative AI development centered around health to bring safety and non-race dynamics to the forefront of AGI development while providing uncontroversially beneficial products and services.

At least a dozen or so additional projects are in the list, but currently lack clear leadership or an institutional home.

Next Steps

Any project in this space requires (i) a worthwhile idea; (ii) leadership; (iii) personnel; (iv) funding; and (v) an institutional home. This list and others demonstrates that (i) is not a major bottleneck. Given other ingredients, FLI's experience is that there are also many personnel available unless a particularly unusual or highly technical background is necessary – for example FLI maintains a list of over 1,000 prospective volunteers. Numerous institutions now exist in the relevant space. Funding is always scarce relative to need but less of a constraint in this space than many others in the nonprofit/civil society sphere. It thus appears that capable, creative, entrepreneurial leadership for projects is currently a major bottleneck, suggesting that the community consider how additional capacity might be developed.

AI Safety Communication: Stakeholders, Narratives, Projects

Session chaired by Jingying Yang, Tasha McCauley, Allison Duettmann, summarized by Jingying Yang

Overview

Reframing AI development in cooperative terms to encourage cooperation towards AGI within and outside of the AI safety community is a worthwhile undertaking with many benefits.

Because AGI does not yet exist, there is currently an opportunity and responsibility to carefully shape these early exposures and in some cases, reshape early impressions before harmful memes take root more permanently. Realistic positive AI narratives can be an effective way to achieve more emotional and human-centered goals such as inspiring cooperation, reducing fear, and inciting action towards positive goals.

Next Steps

Message, medium, and target audience will vary according to specific goals. Below are two ongoing projects to shape AGI narratives that show how message, medium, and target audience can vary depending on the goal of communication. Both projects are open to collaboration with the respective organizations.

The Partnership on AI's Positive Futures from AI project:

- **Motivation:** The pervasiveness of the Terminator/Skynet idea in AI-related journalism and the common dystopian perspective in movies and TV shows about AI makes general audiences who are less familiar with AI safety afraid of less likely aspects of an AI future, such as a nefarious AI agent that plots to destroy humanity. These fears take up emotional space that could be better used to grapple with more likely and nuanced risks on the

path towards building an AI-filled future, such as human enfeeblement, erosion of community if AI is used to create ersatz personal universes instead of enabling deeper connection, etc. This fear can also blind people to the positive possibilities that AI technology can bring which can actually be actualized if we all start working towards these positive outcomes now.

- **Goal:** To improve the state of public discourse by displacing the Skynet narrative and giving people an inspiring, concrete, and technologically attainable alternative positive narrative for society that involves AI.
- **Audience:** General public who does not know much about AI.
- **Message:** This is the core of the project right now. PAI aims to inclusively craft a vision for what positive futures from AI can look like through a few different methodologies. At its core, PAI aims to create a reliable methodology to help people translate their desires, needs, and motivations into a vision for the future in order to make the process of designing the future of society more inclusive. 1) Convene diverse sets of experts to go through a structured set of exercises aiming to elicit fruitful brainstorming about positive futures. The benefit of these convenings is the expertise that can make these visions verifiably technologically attainable. The downside is that smaller convenings can only capture a more limited set of perspectives. 2) Public call for visions. Once PAI develops a reliable methodology that works across cultures to help people visualize the future, PAI will aim to gather as many stories as possible from members of the public, and then aim to analyze these in order to find overlaps between them. Then, PAI will harmonize these visions of the future to find narratives that are commonly appealing and these will be the message to transmit.
- **Medium:** Movies, TV, sci-fi novels, mainstream journalism.

Foresight Institute's Existential Hope primer for the next generation

- **Motivation:** Many young adults who will soon shape our future are plagued by information overload, disillusionment, or disinterest. Similar to how Nick Bostrom's "Superintelligence" sparked the current risk-focus in the AI community, this project provides positive concrete alternatives that talent can get behind to cooperate on.
- **Goal:** Create an onboarding document for the next generation that gets every smart and committed 20 year-old to have the common map of what's at stake for civilization, wrapped in a shared experience that allows them to swap their existential angst for existential hope. Apart from sparking a memetic shift, the primer will provide shortcuts to high-impact career paths, organizations, and communities to join.
- **Audience:** Precocious youth and college students who are aware of a multitude of systems which should be torn down but have very few positive comprehensive visions to work towards.
- **Message:** This will be the book you wish you could send back in time to your younger self: a primer illuminating the promise and peril lying ahead of our civilization. After validating reader cynicism by outlining coordination failures and short termism the book will argue that history has started again and frame our position as in between a big history and a big future. Different chapters will highlight promising trajectories through nanotechnology, biotechnology, intelligence, cyberspace, biosphere, and space. The primer concludes with a call to action for how to realize those promising trajectories.
- **Medium:** A 100-page primer encoding a map of possible positive futures as synthesized from leading thinkers. The book is built in a collaborative process by interviewing ambitious individuals in crucial science and technology domains and distilling their models into action plans.

A Framework for AI Fiduciary Agents

Session chaired by Gaia Dempsey and summarized by Jim O'Neill

Overview

Humanity is a machine created to make plumbing for kings. Therefore, people don't act in their own interest and they make even worse fiduciaries for others. Algorithms have the potential to be more loyal and diligent extensions of the individual's will. They could guess what we would do if we had the time.

Benefits & Advantages

Good information fiduciaries should make decisions in our absence while protecting our privacy. They should be accountable and transparent to us and control digital sharing. One good privacy-protecting example is Almond. If properly designed, these agents could be much more reliable than human fiduciaries. They could decentralize AI power, protect secure enclaves around neural interfaces, and enable a legal construct between AI agents.

Risks & Obstacles

Preserving privacy will limit the training data for machine learning compared to centralized systems, but if AI assistants become ubiquitous without these safeguards, people will suffer. We don't have secure email because government doesn't allow it. Will government allow secure agents?

Next Steps

We should study the literature of principal-agent problems and the analysis of the impact of automation on human performance of Chris Wickens' Human Performance Consequences of Stages and Levels of Automation. To begin to create these agents, we should arrange a hackathon to develop prototypes. We need to ask if an agent can really make decisions in your interest without informing you, and start to decide in which situations you want it to consult you.

5. Reframing AI Trajectories in Cooperative Terms

Intelligent Voluntary Cooperation

Session chaired by Christine Peterson, Mark S. Miller, Allison Duettmann, summarized by Allison Duettmann

Overview

The session discussed a book draft by the session chairs on the topic of Intelligent Voluntary Cooperation: As technology advances, it both increases the potential for automated violence while also bringing new tools that allow us to cooperate ever more richly. The long-term trajectory of civilization will depend on how we navigate this bottleneck. We need to explore decentralized encrypted monitoring layers to prevent violence, and implement emerging cryptocommerce tools to reap the benefits of cooperation. If we want to survive and thrive in an increasingly intelligent future, this fabric must include software intelligences.

To unlock voluntary cooperation with other intelligences, we must first ensure that interactions among humans and those intelligences remain voluntary. Since most of our interactions will be digital, this requires bottom-up computer security. Once we have secured a voluntary base layer, we can open up opportunities for cooperation. We review how software intelligences will cooperate with human intelligences in a voluntary architecture if there is a large diversity of such agents, pursuing a variety of goals that can best be pursued via mutually beneficial cooperation.

Similar to Comprehensive AI Services (CAIS), Intelligent Voluntary Cooperation (IVC) seeks to prevent a unitary takeover of a singleton AI in favor of economic cooperation amongst a diversity of intelligences. While CAIS treats those intelligences as non-agential services, IVC prepares for the case in which we cannot prevent services from becoming agential.

Benefits & Advantages

IVC seeks to prevent a unitary take-over by one intelligent agent in favor of a long-term multipolar world in which a diversity of intelligent agents cooperate. It understands civilization already as a superintelligence that serves human interests. Civilization is superintelligent in the sense that it is vastly better at problem-solving than any individual intelligence. It serves human interests, not in the sense that it "wants anything"; it has no utility function, but it does have a tropism—it tends to grow in certain directions. To the extent that its dominant dynamic emerges from noncoercive voluntary interactions, it is already shaped by human values. It tends, imperfectly, to climb Pareto-preferred paths. By integrating emergent software intelligences into this cooperative framework we can sidestep hard ethical considerations about what it means to align software entities with human values. It is unlikely that we will build a non-human intelligence that is better aligned with human interests than this highly evolved emergent dynamic.

Risks & Obstacles

Any attempts to centralize intelligence capability that upends the multipolar power balance of civilization creates first strike instabilities which could cause nuclear winter by other actors preventing the take over. To avoid this, we should strengthen AI methods that decentralize capabilities toward a variety of specialized goals, and look to existing compensating dynamics to decentralize power, e.g. the US constitution, international mutual defense pacts, and cross-jurisdictional cooperative arrangements across organizations.

Next Steps

The IVC Model will be discussed in detail in an upcoming book on Intelligent Voluntary Cooperation, co-authored by Allison Duettmann, Mark S. Miller, Christine Peterson.

Policy for Decentralized Take-Off Scenarios

Session chaired by Robin Hanson, summarized by Allison Duettmann

Overview

If successful, a decentralized AI take-off scenario may allow side-stepping some of the risks of a singleton AGI scenario by establishing a framework in which a multiplicity of AI entities co-exists. Here we address some novel concerns that may arise under a decentralized take-off scenario.

The following is an incomprehensive list of question clusters, ranked according to the urgency assigned by the group:

- Small kills all: Does the theoretically plausible trend of increased vulnerability of the world actually hold empirically in current global developments? Can we rely on examining past trends for extrapolating future trends? For instance, if medieval societies spent a larger fraction of their economies on offense than defense, could this count as data point to counter claims that the global offense vs. defense balance is worsening?
- Low human influence: Will humans decline in influence relative to a future that includes decentralized instantiations of artificial intelligences? Will future intelligences share human values and aesthetics? Can we influence future variance in AI values by programming them to be complementary to human environment and

human wellness? Will AI descendants count as human enough for the purpose of mattering?

- Conflict equilibria: Can we reach peaceful global power equilibria or will a future world become Medieval-like with conflict as a continuous equilibrium?
- Larger generational gaps and loss of consciousness: Is it possible that competition causes a race to the bottom that, over multiple generations, may lead to loss of human value up to the destruction of consciousness?
- Pollution of information environment: Will AI make it harder for humans to decide what to believe, leading to an impaired ability to do long-term planning?
- Will decentralized AIs worsen or improve the depletion of resources?
- How can we manage the transition to an increasingly automated world? Is inequality bound in a world of labor, but not in a world in which labor becomes less important as work gets automated? How do proposals to manage the transition compare? For instance, how does a Windfall Clause by which companies precommit to taxation above a certain level of profit to be shared across the population compare to more standard economic solution to facing risks, e.g. purchasing insurance against job automation? How would insurance solution work? For instance, if insurance companies buy stock in individual AI companies, how can we ensure that insurances have global reinsurance to avoid secondary risk of one's local base insurance not actually covering the risk?

Next Steps

When considering how a future world of decentralized AI may differ from the current world, a systematic survey would be useful to determine wider root causes of the question clusters above. Further research may reveal whether the concerns about decentralized take-off scenarios above are exacerbated by AI or whether those are general concerns about the past, present, and future of civilization, independent of their relationship to AI. For instance, war is a problem that is a prevalent concern about decentralized AI-futures but many of the future warfare problems related to intelligence are some which are already possible with current technologies like automated weapons or automated cyber-attacks. It is plausible that global destruction is something that actors are becoming increasingly able to do, but if future problems bear some resemblance to current and past problems, we may look to history for guidance. In addition, further research into the dependencies of the long-term race between offense and defense dynamics could reveal whether useful analogies to past multipolar global situations can be established.

Conclusion

This report summarizes ongoing work and desired future work on AI cooperation via five clusters:

1. The need for cooperation on AI geopolitics;
2. Crucial theoretical considerations regarding AI cooperation;
3. Short-term AI needs and applications as grounds for cooperation;
4. Concrete ongoing projects seeking to advance cooperation; and
5. Reframing AI trajectories in cooperative terms.

In sum, we need to get better at cooperation to navigate the emerging geopolitical landscape around AI. In addition to working on better theory and models around AI cooperation, this involves work on the beneficial short term applications of AI, projects with the distinct focus of motivating cooperation across crucial stakeholders, and non-adversarial framings of AI narratives and trajectories. We point to encouraging ongoing work in those areas with the hope to encourage future research.

FORESIGHT INSTITUTE

Artificial General Intelligence: Toward Cooperation