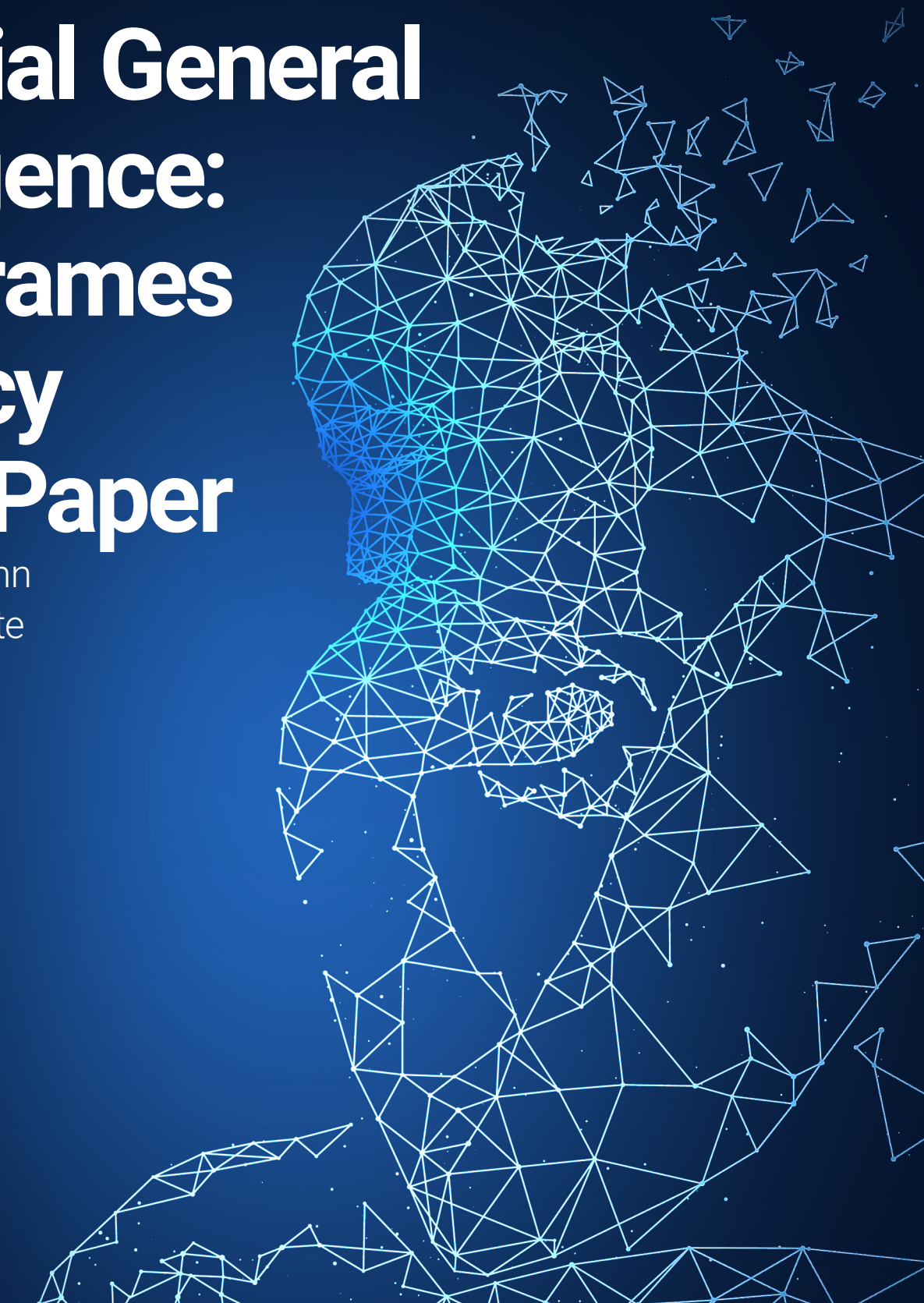
FORESIGHT INSTITUTE

Artificial General Intelligence: Timeframes & Policy White Paper

Allison Duettmann
Foresight Institute



Artificial General Intelligence: Timeframes & Policy White Paper

Allison Duettmann, a.duettmann@foresight.org

Based on AGI Timeframes & Policy Workshop

Hosted by Foresight Institute in San Francisco on August 10, 2017

Foresight Institute is a non-profit organization focused on supporting the upsides and avoiding the downsides of technologies of fundamental importance for the human future, especially molecular machine nanotechnology, cybersecurity, and artificial intelligence. Aside from awarding the Feynman Prize, and granting the Foresight Fellowship, the institute hosts invitational workshops at the intersection of scientific disciplines, public salons on societal topics, and produces in-house research focused on policy recommendations for science and technology.

This meeting was initiated by the observation that some researchers' timelines for AGI arrival were shortening and the perceived increased urgency for drafting potential policy responses to the related arrival scenarios. This report outlines participants' timeline estimates for the achievement of Artificial General Intelligence and problems associated with arriving at timeline estimates. Rather than focusing on investigating exact timelines in more detail, it is more instructive to consider different high risk scenarios caused by Artificial Intelligence. The main part of the report focuses on three high-risk scenarios, (1) cyber security, (2) near-term AI concerns, and (3) cooperation leading up to Artificial General Intelligence. While some immediate recommendations for further investigation of potential policy responses were made, the meeting's main intention was not to reach consensus on specific topics but to open up much-needed dialogue and avenues for cooperation on topics of high importance for policy considerations pertaining to Artificial Intelligence.

Publication date November 2017

This report can be referenced as follows: Duettmann, A. (2017). "Artificial General Intelligence: Timelines & Policy White Paper." *Foresight Institute*. Available at foresight.org

Participants

Anthony Aguirre, Future of Life Institute
Shahar Avin, Centre for the Study of Existential Risk
Haydn Belfield, Centre for the Study of Existential Risk
Malo Bourgon, Machine Intelligence Research Institute
Niel Bowerman, Future of Humanity Institute
Tom Brown, Google Brain
Miles Brundage, Future of Humanity Institute
Allison Duettmann, Foresight Institute
Peter Eckersley, Electronic Frontier Foundation
Carrick Flynn, Future of Humanity Institute
Katja Grace, AI Impacts
Melody Guan, Google Brain
Brewster Kahle, Internet Archive
Durk Kingma, OpenAI
Tom Kalil, UC Berkeley
Richard Mallah, Future of Life Institute
Mark S. Miller, Foresight Institute
Mark Nitzberg, Center for Human Compatible AI
Jim O'Neill, former US Dept. of Health and Human Services
Christine Peterson, Foresight Institute
John Salvatier, AI Impacts
Andrew Snyder-Beattie, Future of Humanity Institute

Table of Contents

AGI Definition

The proposed starting definition of Artificial General Intelligence (AGI) for the meeting was Nick Bostrom's definition of superintelligence as "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest" (Bostrom, 2014). Some participants pointed out that it is important to not conflate the terms "Artificial General Intelligence" and "superintelligence", referring to literature that aims at a more precise explanation of the harder-to-define term of "AGI" (Muehlhauser, 2013). Some participants used alternative definitions, e.g., a more inclusive definition of superintelligence, according to which the whole of civilization is the relevant superintelligence, composed of both human and machine intelligences (Peterson, Miller, Duettmann, 2017). A related definition of Artificial Intelligence suggests that corporations can already be classified as first generation AIs (Kahle, 2014). Generally, human-level intelligence might not be a good standard for general intelligence: It is not general, due to the hardware constraints of the human brain, as well as evolutionary artifacts and biases that constrain human thinking.

Timelines

Timeline skepticism

While having better AGI timeline estimates would be useful, there was considerable skepticism toward the ability to form informative estimates for AGI arrival based on current information. Reasons for skepticism were focused on concerns about how individuals form timeline predictions, and how group dynamics can influence an individual's abilities to make estimates:

Individual biases that can decrease the individual's ability to predict

- Humans are inherently poor at estimating the likelihood of novel events because they tend to base estimates on stylized facts about the way the world works and are prone to several cognitive biases (Wikipedia's list of cognitive biases). E.g., the availability bias might lead humans to underestimate the speed of AGI advancement, and other biases, e.g., the planning fallacy, might lead humans to overestimate the speed of AGI advancement. A growing body of research shows that experts are not immune to a variety of biases when making predictions, which suggests that expert surveys of timeline estimates can potentially be misleading in their authoritative force (Armstrong, Sotala, 2012).

Some strategies to optimize predictions

- Improve one's forecasting abilities via tools described in Superforecasting (Tetlock, 2016), or via projects like the Good Judgment Project
- Calibrate one's credence level to more closely reflect past success rate (e.g., Credence Calibration Game)
- Get feedback on how well one's predictions describe real events by joining prediction markets (e.g. Metaculus). While predictions exist even for long-term events (e.g. Long Bets), it is hard to use them as

training data, because feedback on whether or not a given prediction is correct is delayed. Predicting precursor events that would make the long-term outcome more likely could be an alternative here

- Regardless of the above difficulties, there are some strategies to extract value from expert forecasts (e.g., judgmental bootstrapping), and a recent expert survey to forecast AI advancement has made a significant effort to avoid known biases (Grace, Salvatier, Dafoe, Zhang, Evans, 2017)
- For a good resource on strategies for optimizing one's predictions, see forecastingprinciples.com

Group dynamics that can decrease individual's ability to predict

- Echo-chamber effects within an intellectual peer group leading to double-counting of one person's position (i.e., a position is not traced back to its origin but spreads as rumor through the community, further increasing its impact on other individuals' positions)
- The result is often groupthink that further exacerbates individual's bounded abilities to form reasoned predictions

Some strategies to counter harmful group dynamics

- Taking non-mainstream views in the expert community seriously to counter the artificial agreement on some mainstream view caused by groupthink
- Relying on independent sources of information and seeking out information channels that are contrary to those of one's peer group to decrease the echo-chamber effect
- Give more weight to views of individuals with structured values and models for tracking the success of their predictions
- Timeline specific: When experts views differ greatly, a broad distribution of timelines, rather than insisting on one consensus timeline, appears more useful. Nevertheless, analyzing the correlations between experts' views can be informative in some cases (Muehlhauser, 2016)

A further point leading to timeline skepticism is not based on an epistemic concern but instead on a motivational concern: Timelines are not exogenous to the behavior of relevant actors but can be influenced, e.g., by capability research, safety research, increased coordination among actors, political events, etc. Stipulating techno-deterministic timelines might evoke a harmful deterministic perception of the arrival of AGI that could lead to relevant actors giving into fatalism.

Timeline findings

Regardless of the group's general skepticism on the accuracy of AGI time estimates, some noteworthy findings emerged:

- Given timeline estimates varied widely, from 0–80 years, with 0 years referring to either (a) superintelligence in the form of civilization already existing, or (b) some primitive, non-ambitious AGI in the working definition already existing
- Nevertheless, participants' probability distributions for different timeframes showed a trend toward shorter timeframes. While participants gave AGI arrival of >25 years considerable weight in their probability

distribution, many participants' probability distributions had significant mass at AGI arrivals of <25 years, and there was a non-negligible consideration of extremely early AGI arrival between 5–10 years.

■ Seed AGI (an AGI that improves itself by recursively rewriting its own source code without human intervention) was considered as a likely scenario for AGI arrival. This type of AGI could yield especially short timelines, potentially between 10 and 25 years, due to its initial primitive character

Need for models, and how different models affect one's estimates

It was suggested that one way to improve one's estimates for AGI timelines is via the use of models. However, depending on the likelihood that one assigns to different factors in one's model, AGI timelines will vary greatly.

Potential factors and their influence on models

■ Universal quantifier vs. hard-coded infants: Only one of the multiple approaches to AGI has to work in order for us to reach AGI. This combined with the rigor with which AI is currently pursued may make a shorter timeline more likely. Alternatively, one could believe that there is some inherent hardwiring in biological creatures that allows them to learn certain tasks quickly (e.g., giraffes are able to walk shortly after their birth), which would have to be computationally implemented in an AI, and may make a later timeline more likely. By assigning a higher probability to different factors, one might arrive at an earlier or later timeline.

■ Importance of software vs. hardware: Timelines will vary whether one tackles AI from a software approach (Pieter Abbeel's approach to robotics that relies on Sensorimotor Deep Learning), whether one thinks that the limitations in current hardware cannot yet be sufficiently compensated for by software tools (Goertzel, 2012), or whether one thinks that entirely novel forms in hardware, e.g., optical computing, are required to achieve AGI.

■ Researchers might analogize success in AGI to success in other difficult scientific problems. For example one might compare it to:

- Quantum gravity: a very hard, quite old problem that could well take many decades more to solve, but is primarily limited by missing insights, so there could be a breakthrough on shorter timescale.
- Building a cell from scratch: we understand the problem well enough to know that it will be many, many decades before we can do it, because the task is so complex and demanding that it's far beyond current technology and abilities.
- Quantum computers: a very difficult technical task, but it is plausible that continued refinement and improvement of current techniques, along with plausible development of new related ones, could lead to success within several decades.
- Many researchers have generally considered AI as combining the difficulties of quantum gravity and cell-engineering. This may still be the case, but recently researchers are according more probability to it being like quantum computers, which is quite a different picture.

A general, but simple rule of thumb that should be reflected in models for AGI arrival times is that improvements and increases in four important factors—algorithms, hardware, human skill, and time—make early arrival dates much more likely.

Milestones and metrics to assess AGI development

Apart from models, milestones can be helpful to assess progress in AGI development. These milestones should be general, rather than narrow: Measuring progress in individual domains, e.g., Natural Language Processing, is not a good indicator for progress on AGI, because a seed AGI could be very limited in those domains initially but learn fast. So while it would initially perform worse on individual domains than a narrow AI which is optimized to perform well on the given milestones, it could be the best contender for AGI. This reasoning is related to Goodhart's Law, that states that when a measure becomes a target, it ceases to be a good measure.

Potential strategies to generate milestones and metrics

- Model milestones according to milestones used in related domains, e.g. comparative psychology
- Measure by how much AI reduces the time to the next scientific breakthrough and use this as metric for progress
- Make a comprehensive analysis of the problem space that needs to be solved for AGI combined with an analysis of the state-of-the-art. An example is the Electronic Frontier Foundation's recent project that divided the problem space into solved problems, problems that are solvable in 5 years, and problems that are hard to solve soon (Eckersley, Nasser, 2017). According to the analysis, none of the problems in AGI were seen as in principle too hard to solve. An advantage of this approach is that one could compare the number of problems remaining to solve AGI with the number of problems remaining to solve AGI safety, to encourage work on AGI safety (if a similar model was developed for AGI safety). A problem with such an approach is that it might underestimate the speed of AGI progress, as it neglects the possibility of potential breakthroughs which could dramatically speed up progress on problems. To counter this, one could support the analysis of remaining problems with an analysis of the space of potential breakthroughs and their effect on the speed of problem solving.
- Survey the main AI safety projects and their strategies to gain more clarity of the general problem space, similar to the Survey of Artificial General Intelligence Projects for Ethics, Risk, Policy that was recently released (Baum, 2017).

AI Safety approaches should be robust across a broad range of timelines

Given the difficulties in reliably estimating exact timelines for AGI, and given the great time differences between estimates amongst experts, AI safety approaches should be robust across a variety of timelines, rather than focusing on a specific AGI arrival scenario. Generally, approaches that are robust across a broad spectrum of risks are to be preferred to approaches that are risk-specific. However, on the margin it makes more sense to shift effort to work that's relevant on shorter timelines since that work would be especially urgent if we are in a world where timelines are short. Especially if that work is also likely to be helpful in worlds with longer timelines. Some safety approaches that were discussed at the meeting have contradictory recommendations, e.g., on whether to speed up AGI development or delay it.

In favor of speeding up AGI development

One safety strategy might recommend to speed up AGI development because delaying AGI could be dangerous for several reasons, for example:

■ A later AGI achievement means that the hardware that is available at the point of achievement is likely more powerful than at the point of an early AGI achievement. Powerful hardware makes a hard take-off more likely, by which an AI rapidly self-improves and could achieve superintelligence in the number in weeks, days, or even hours after the last human intervention. A hard take-off is dangerous as it allows little to no human interference in the type of AGI that is evolving during the rapid self-improvement period, making it less likely that the AGI will uphold human values (Bostrom, 2014).

■ Another point in favor of an early creation of AGI is bridging the acute-risk period that humanity is in as quickly as possible. We are currently in a state risk where anyone who can develop an AGI, could deploy it and potentially cause a catastrophe. We will continue to be exposed to this state risk until that situation changes. To transition out of this state risk, one could attempt to create a minimum viable AI system that would have a transformative impact on the world such that a takeover by a hostile party or AI might be prevented. One could focus on refining and upgrading this minimum viable AI system after the period of a hostile party potentially winning the race is bridged. This strategy would eliminate state-risk, but introduce transition risk, a risk that arises when transitioning from one state to another: Rather than remaining in the state where anyone could deploy an AGI and cause a catastrophe, it involves deploying an AGI, which could potentially be a risk but only at the time of deployment. If the deployment is a success (i.e, didn't cause a catastrophe), the risk wouldn't be present anymore.

In favor of delaying AGI development (possibly indefinitely)

An alternative strategy (Drexler, 2017) recommends pursuing non-agent-centered approaches to providing general AI services, potentially postponing the development of AGI until the safety problem is solved. This approach is based on the observation that AI is advancing through distributed research and development, and that R&D tasks can be incrementally automated as AI advances. Looking forward along this path, rapid AI-enabled AI-technology improvement does not require engineering self-improving agents (putting the whole R&D chain into a conceptually-opaque box). This approach scales to superintelligent-level AI technologies, and development services that operate at this level can presumably implement any desired AI service. In other words, comprehensive AI services do not require autonomous agents with universal capabilities.

The R&D-automation/AI-services model suggests that the marginal instrumental value of AGI may be low, and that we should expect AGI to be developed relatively late and in the context of a world with access to more tractable high-level AI resources. If so, then high-level AI resources could be applied to the problems of managing AGI. This situation is different from what is usually envisioned, in which an AGI agent embodies a dominant share of the world's intelligent resources.

Regardless of whether one estimates or favors short or long timelines of AGI development, AI safety demands immediate and concerted effort due to the complexity of the problem. Solving AI safety potentially involves solving difficult subproblems in areas as diverse as ethics, technical alignment, cybersecurity, and political coordination. Since we have no confidence on how fast we can solve these problems, we should not take later arrival scenarios as reason for complacency.

Three Risk Scenarios

Catastrophic and existential risk are the relevant concern

The growing focus on AGI safety results from the concern regarding risks that can be catastrophic or existential to current and future potential life. However, insofar as the main underlying concern is risk, it is instructive to consider related areas of risk. Even shorter-term risks arising from narrow AI merit concern, e.g., even without general intelligence a software system can take advantage of big data and cause potentially catastrophic consequences (more on this topic below). Rather than focusing on investigating exact timelines for AGI in more detail, different high-risk scenarios related to AI were investigated. The scenarios range from (1) current risk due to poor cybersecurity, to (2) extremely short-term risk from near-term AI applications, to (3) short or long-term risk arising from AGI, especially cooperation.

I Cybersecurity flaws
II Near-term AI issues
III AGI concerns

I. Cybersecurity flaws

Cybersecurity merits its own scenario analysis because (a) it may be an important but undervalued prerequisite for reliable AGI and (b) it is potentially useful to avoid other risks that are themselves of catastrophic or existential nature.

Cybersecurity as desirable for expecting an AGI to execute instructions

Cybersecurity is desirable for expecting an AGI to reliably execute instructions (Peterson, Miller, Duettmann). Even if the AI alignment problem, which is one of the hardest problems to solve in AI safety (Soares, 2017), was solved, we could not necessarily expect the AGI to reliably execute the desired goals without proof that the hardware and software system within which the AGI is built is itself reliable. Current computer systems are vulnerable to an escalating number of attacks and little time passes without novel insecurities being exposed (in 2015, Symantec discovered on average one novel Zero-Day attack per week, twice the rate of the year before). For narrow AIs safety failures are likely to not exceed a moderate level of criticality, however for general AI, a single failure may cause a catastrophic event without a chance for recovery (Yampolskiy, 2016).

Possible failure modes due to lack of provable cybersecurity

- Malicious hackers could access the system to sabotage or reprogram the AGI (Yampolskiy, 2016; Bostrom, 2017). Given the immense advantage that AGI conveys to its owners, the threat of a cyberattack to advanced actors in AI is potentially big. The threat level increases, especially with the rise of sophisticated AI attacks that are increasingly more capable of discovering vulnerabilities and inventing attacks themselves, e.g., rather than merely employing Zero-Day Attacks, future AI-powered software should be able to analyze the system being attacked and find entirely new, previously unknown Zero-Day Attacks (Peterson, Miller, Duettmann, 2017).

- An unintentional bug could disturb the AGI's goal execution. While this problem may seem small in our current narrow AI world, given the potential reach of an AGI's actions, even a slight bug could cause catastrophic consequences.

- An AGI could discover and exploit vulnerabilities in its own code, e.g., by reprogramming itself in unintended, dangerous ways, or in the case of a confined AGI, by breaking out or gaining access to the internet.

Cybersecurity flaws as catastrophic risks

Cybersecurity is not only desirable for AGI safety but for the safety of most systems that rely on computational systems. In our current world, this is much of the societal ecosystem (Peterson, 2017).

Possible catastrophic risks related to cyber security, that are unrelated to AGI risk

- Attacks on the vulnerable electric grid, leading to power outages across connected parts of the US (Peterson, 2017). A persistent electricity outage could not only lead to shortages in the food supply but might reduce overall coordination of humans such that the carrying capacity for the planet could be reduced. A study by University of Cambridge and Lloyd's estimated that damage from one regional attack could exceed $1 trillion (Rashid, 2015).

- Attacks in the near future that allow bad actors to remotely steer self-driving cars or drones, causing accidents of inconceivable scales, or turning UAVs into weapons of mass destruction.

The problem of cybersecurity vulnerability is already severe but could worsen dramatically in the near future, if the offense progresses parallel to progress in AI, while the defense advances slower or stagnates. The problem with respect to defense is that a multi-trillion-dollar ecosystem is already built on the current insecurable cyber foundations. The likelihood of getting political adoption of defense approaches that require rebuilding parts of, or the entire ecosystem, from scratch are low, especially after the underwhelming responses to recent cyber attacks that

reveal massive vulnerabilities (Miller, Peterson, Duettmann, 2017). Nevertheless, there are cyber defense approaches that can help tip the cyber world toward favoring defense over offense.

Possible cyber defense approaches

- Incentive change: Incentivize those hackers currently working in favor of offense to disclose discovered vulnerabilities, so that defense strategies can be developed.

- Defense contest: Incentivize hackers to work toward defense via contests that are designed to work on defense strategies. These contests would be different from previous DARPA's Cyber Grand Challenge, is as they tended to encourage work on the offense (Gillula, Cardozo, Eckersley, 2016).

- Policy to expose vulnerabilities: The U.S. federal government already collects a large body of knowledge on existing vulnerabilities via the National Security Agency (NSA), making the NSA the potentially greatest red team known to man. The NSA could privately disclose all vulnerabilities it collected of actors to these specific actors in the system (vendors and customers), together with an official deadline that all vulnerabilities that are still existing in a given period of time would be publicly revealed. This would make every actor who refuses to rewrite their security infrastrastructure before the deadline vulnerable to attack and could present enough of an incentive for the actors to fix their vulnerabilities.

- Employ existing safety strategies: Implement existing safety strategies that are known to work. It is widely believed that improvements to safety are a matter of technological discovery or need for new research. However, most of the techniques required to build systems that are largely secured from cyberattack, with a few exceptions, have already been known since the 1960s and 1970s, e.g., capability-based security (Dennis, Van Horn 1966). An additional approach is via the use of verified software, such as seL4. The seL4 microkernel is our best example of an operating system kernel 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 (Peterson, Miller, Duettmann, 2017). These techniques could be adequate for defense if society could somehow reconstruct the computational world, from its beginning, on top of those techniques.

- One scenario on AI-powered cyberattacks that was discussed at a recent meeting of AI experts on Envisioning and Addressing Adverse AI Outcomes suggested that AI could significantly exacerbate the existing cybersecurity threat, and suggested a current lack of defense strategies (Bass, 2017). A future R&D research project could test whether AI can speed up security development in fields such as automated proof generation or red team attacks. However, in previous cases of AI applications for security enhancement, specifying the problem such that it could be adequately interpreted by AI has proven challenging.

- Using the blockchain ecosystem as role model to remodel current societal ecosystem: One reason the mainstream adoption of secure computing is delayed is that the overall software ecosystem is not currently "hostile enough," i.e., companies and institutions can be successful even though they implement their systems in architectures that are insecurable. A counterexample to this is Ethereum's current approach. Both Bitcoin and Ethereum are evolving in an ecosystem that is already under very hostile attack pressures, with the 2016 DAO exploit being a case in point. When insecurity leads to losses, the players have no recourse for net compensation; losses are real. Such systems that are not bulletproof will be killed early and visibly, and therefore these ecosystems remain populated only by (so far) 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 can spread quickly enough throughout

today's computing ecosystem, then a successful genetic takeover scenario from the current insecure system to a blockchain-based system might be achieved (Peterson, Miller, Duettmann, 2017).

Objections to increasing cybersecurity

One objection to increasing cybersecurity defense is that by making cyber attack an easy, profitable attack vector, kinetic warfare could be avoided, which is likely more harmful. If an actor has a considerable advantage in AI development and seeks to hinder a second player from catching up, it could use its increasingly sophisticated AI to launch a cyber attack to eliminate the runner-up's capabilities, rather than resorting to kinetic warfare with that player. Thus, cyber vulnerability might allow one actor to non-violently slow down another actor. If non-kinetic warfare results in fewer casualties, it is preferred to kinetic warfare, so leaving the option of cyber attack as an attack vector could potentially be beneficial. An organization that is leading AGI development could potentially use advanced AI techniques to launch cyberattacks on runners-up to gain more breathing space to work on AI safety to develop a safer take-off to AGI.

Some participants objected to this point that even if non-kinetic warfare is superior to kinetic warfare, a stronger AGI actor hacking the runner-up could potentially be destabilizing to the whole ecosystem in its own right. There was no consensus on whether or not this scenario is more stable or unstable than the available alternatives.

II. Near-term AI issues

Near-term AI issues are composed of concerns regarding "narrow AI": AI that is not general in its intelligence (yet) but that is trained to perform one or more tasks very well. Five major areas of concerns were discussed:

1. Autonomous weapons
2. Corporate concentration of power
3. Phase change in science and engineering
4. Surveillance, control, social engineering
5. Major unrest from economic disparity

1. Autonomous Weapons

Autonomous Weapons that are equipped with artificial intelligence could lead to an increase in catastrophic and existential risks, for at least three reasons: Bias in favor of military offense over defense, a power shift amongst military powers, and lack of accountability:

- Bias in favor of military offense over defense: One conceivable threat enhanced by near-term AI advances would be Unmanned Aerial Systems (UAS) equipped with AI components, e.g., AI-powered facial recognition, manned with Improvised Explosive Devices (IEDs) and abused as improvised guided weapon systems by terrorists. While the current size of the explosive device that can be deployed by a drone is relatively small, the ability to deploy such weapons to a specific, vulnerable location, at speed and with accuracy, makes them a significant threat (Davies, 2017). Such weapons systems equipped with AI-powered facial recognition would greatly increase targetability of offense. Given that AI-powered drones could allow a small number of actors to cause more harm, they could shift power balances dangerously in favor of offense.

- Power shift amongst military powers: Apart from advantaging offense over defense, rapid advances in Autonomous Weapons can also shift the power balances among different players, upsetting existing military power balances. In response to the advances in autonomous weapons, the Convention on Certain Conventional Weapons agreed in 2016 to formalize their deliberations on autonomous weapons and form a

group of governmental experts—90 countries are expected to participate in this year's meeting at the UN in Geneva.

- Lack of accountability: Since Autonomous Weapons are often steered remotely, if at all, attribution of the aggressor of an attack launched by such a weapon can be difficult. This lack of attribution could give rise to false flag attacks, with attackers pretending to be certain other actors to jeopardize the relationships between target actors. This would further decrease international stability and could lead to a lower threshold for war.

Generally, the catastrophic risk community should take the threat of Autonomous Weapons more seriously, because these weapons will likely get smaller, less expensive, and more widespread. The cybersecurity threat of Scenario 1 above is relevant for the Autonomous Weapon consideration also, as cybersecurity attacks could enable bad actors to turn generic remotely-controlled systems into Autonomous Weapons. Campaigns against Autonomous Weapons include advocacy efforts by autonomousweapons.org, and the Future of Life Institute's Open Letter from AI & Robotics Researchers.

2. Corporate Concentration of Power

AI techniques that allow individualized targeting of consumers (e.g., Natural Language Processing or Sentiment Analysis) could be used as super-propaganda tools that greatly increases the power of individual corporations. The Cambridge Analytica role in the recent US election is a case in point for how easily individuals can be analyzed and influenced according to their online behavioral data with far-reaching real-world consequences. As corporations grow, their access to individuals' data increases, resulting in better calibrated AI-propaganda tools that further cement the power of a few large multinational corporations. Via this feedback loop, wealth disparities could increase, not only among individuals but also among corporations, leading to large Gini coefficients among corporations.

As some corporations grow, their power relative to nation-states is likely to change as well. Whether the recent Apple vs. FBI encryption dispute was public theater or real, the possibility of a corporation publicly countering the government on this scale signals an increase in power of corporations. As corporations increase in size, there may be a threshold at which, from a practical perspective, they become too large to refuse giving out information to the government; once a company is very large, it cannot simply shut down to avoid complying with a government order, an option that is open to small companies such as Lavabit, which chose to suspend its email service rather than submit to a court order requiring surveillance of customers. However, generally, the power of corporations is likely to increase via AI, and as some corporations equal or surpass some nation-states financially and get state-level influence, governments could become comparatively less powerful.

While concentration of power could be more risky in states than in corporations, the incentives of corporations can be problematic as well, as seen in the case of Exxon and the tobacco industry. To unmask the mixed bag of corporations' incentives and increase accountability, non-governmental organizations such as the Electronic Frontier Foundation and similar groups could push for increased transparency of corporate processes and operations.

3. Phase Change in Science & Engineering

Currently, most discoveries in science that involve AI are made by a human-machine hybrid, rather than AI alone. However, there are some promising findings for AI-steered research, e.g., the long-unsolved problem of regeneration of flatworms was solved via an ML algorithm in under 42 hours (Lobo, Levin 2015), and there is increased efforts to create AI systems that can propose hypotheses for testing. As AI-powered science and technology R&D increases the rate of discoveries, it has the potential to lead to unanticipated risky discoveries of the magnitude of catastrophic,

or perhaps even existential risk. While the potential discovery of easy-to-abuse novel findings, tools, and technologies is a risk that humanity faces generally with the pursuit of science, AI certainly speeds up scientific progress, allowing even less time for already lagging research of the risks of long-term effects of high-impact novel technologies (Snyder-Beattie, Cotton-Barratt, Farquhar, 2016).

AI has the potential to speed up most scientific fields, but AI tools such as improved simulation could particularly benefit material sciences, such as advanced biotechnology, and molecular machine systems, leading to significantly faster advances in these fields. While in the past the resources needed to build nuclear weapons made it almost prohibitively expensive for individual actors to build nuclear weapons, dangerous advanced biotechnology or molecular machine systems applications could in theory be built by a small number of people in a small lab with a small amount of resources (Duettmann, 2017). Ultimately, such developments of risky scientific tools could put the cost associated with harming a great amount of people on a Moore's Law-type curve.

4. Surveillance Control

AI has the potential to facilitate government surveillance, not only in strengthening the existing top-down surveillance apparatus but also via novel strategies, e.g., near-perfect facial and voice recognition.

While one conceivable positive aspect of the increased observation of people's lives could be increased availability of truthful sources about events, an alternative, dangerous scenario is possible as well, aside from the obvious totalitarian risk: As it becomes consistently easier to counterfeit footage or recordings of events , evidence that was once trustworthy becomes "up for hack" (Adler, 2017). If AI enables false evidence for any event to be convincingly fabricated, the claim to truth becomes increasingly harder to prove. Rather than truth itself becoming fractured, there could simply be too much noisy data to locate a true claim at the speed of the media and in a manner that is convincing to a great part of affected actors or the population at large. Countering this trend requires not only to make truth tractable again but also attractive again. Some tools that were proposed include:

- Tools like Hypertext via its mechanism of transclusion could in theory encourage people to be more critical of arguments. However, this presupposes that people form beliefs based on what is true. If people pursue tribal signalling and entertainment instead, a tool like Hypertext could potentially increase people's polarization into clusters of like-minded ideologies.
- "Oracles", i.e. consensus-based truth-determining systems, could be a more promising mechanism to incentivise truth-seeking. Generally, ideology-based predictions get swept out early and monetary incentives to improve on other people's bets are strong incentives to report truthfully. Those oracles are a major goal (and requirement) of blockchain-based prediction markets like Augur or Gnosis.

5. Major Unrest

Both the arrival of AGI and the path to AGI will likely have great allocational consequences for wealth, status, and power. Regardless of whether the results are as severe as a radical concentration or a radical redistribution of income and influence, the changes to wealth, status, and power can be grave enough to lead to friction in the demographic group that is disadvantaged (Bostrom, Dafoe, Flynn, 2016). To avoid major unrest, strategies need to be put into place. Examples that were mentioned include:

- Universal Basic Income (UBI)
- Universal Basic Outcome (UBO)
- Universal Basic Provision (UBP)
- Allocation of unclaimed resources via Inheritance Day (Peterson, Miller, Duettmann, 2017)

III. AGI Concerns: Cooperation and Great Powers

There are different ways to divide up the relevant players in the AI/AGI space, e.g., divisions according to nation-states, type of organization, or ideological closeness. Two great powers that will likely be prominent in most divisions are the US (and like-minded allied countries, e.g., UK, Japan, Israel), and Asian actors, especially China. Establishing coordination on AI strategies among these great powers, is thus a primary concern on the path to AI safety.

Unipolar vs Multipolar

Prior to discussing possible coordination strategies, a more fundamental question concerns whether more or less players would make safe AGI development outcomes more likely. There is considerable disagreement at the meeting on whether a multipolar world, in which multiple players are simultaneously working toward AGI, is a safer world than one in which one player is the only one in a realistic position to reach AGI. Both scenarios have advantages and disadvantages and a thorough analysis of these features is beyond the scope of the report. See below for a primitive outline of some arguments in favor and against both worlds.

Multipolar World

- In favor of a multipolar world: Some arguments in favor of a multipolar world point to historical examples of totalitarian regimes to make the case that extreme concentration of power invites corruption and that a decentralized checks-and-balances system of a multitude of players is safer than trusting a small elite to act in the interest of many. However, not all unipolar worlds are necessarily totalitarian by definition because the amount of latent control a player can enforce is not the same thing as the content of what is being enforced, e.g. one could imagine a totalitarian libertarian regime, that emphatically disallows violence or coercion between individuals, but otherwise does very little.

- Against a multipolar world: The multipolar world needs to ensure stable and continued coordination amongst a multitude of different players who have adversarial incentives to cheat in cooperative situations. A multipolar world might only be safe on the long-run if defense dominates offense (either directly or through offensive "second strike" ability) throughout the technological development trajectories of the multiple poles. If multiple actors have: 1) conflicting goals 2) an occasional small window of offensive tech dominance before it is neutralized by better defenses by others, and 3) it is likely others will gain offensive tech dominance for small windows as well (through the random walk of tech development trajectories) then each state has a strong incentive to "defect" and attack during their window of decisive strategic advantage. Even leaving that aside, it could demand an arms race as each pole must waste resources on defensive or second-strike capabilities.

Unipolar World

- In favor of a unipolar world: Some arguments in favor of a unipolar world argue that the competitiveness in a multipolar world invites to an unsafe race toward AI. The great first-mover-advantage tied to reaching AGI could incentivize players to cut corners on AI safety, while a more unipolar world, in which a single player has considerable advantage over others, allows the leading player to slow down and focus on safety when necessary.

- Against a unipolar world: The unipolar world needs to ensure that one player stays in a positive, significantly advantaged leadership position, which is hard because the incentives for other parties to threaten the hegemony via other threats, e.g., nuclear weapons, are likely strong. This objection assumes that there is no

strong prospect of mutually assured destruction in the case of threatening the leader and b) the singleton is relatively weak in the amount of control it can exercise. Such scenario in which other players are capable of putting up a strong challenge to the leading player might more accurately be described as a "weak" multipolar world with a large amount of cooperation rather than a real unipolar scenario.

There was no consensus regarding the desirability one of the worlds over the other, and an evaluation of the optimal polarity of a world that leads to AI safety requires further investigation.

Cooperation strategies

Whether or not a unipolar or multipolar world is more desirable for AI safety, it is instructive to consider cooperation strategies among a variety of players in a multipolar world, either to cooperate to create a unipolar world or as a safety prerequisite in a multipolar world. Five possible cooperation pathways were briefly discussed:

- Data-trade as naturally increasing cooperation: It is possible that the increasing demand for data by AGI-developing players could incentivize the players to trade data among each other, leading to a more cooperative environment. However, the data available freely on the internet could already constitute a sufficient data-set to create AGI, once the prior to make sense of it is sufficiently developed. Relying on data-trade alone to naturally favor cooperation is not sufficient.

- Classifying permitted AI levels to agree on regulation: One strategy to favor cooperation could regulate the permitted levels of AI for players to slow players down at a critical moment while AI safety strategies are developed. 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 (e.g., green, yellow, or red with descending permissibility) with red 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.

- Establish a global council on regulation of AGI: The idea of creating a global leadership council on AI safety, eg. 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 in a centralized, transparent manner, which is important, given that the AI safety problem consists of ever-growing 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.

- Smart contracts to enforce cooperation: If regulations or other cooperation treaties were developed, one strategy to enforce cooperation with these agreements would be via smart contracts. Because smart contracts allow players to bind themselves to a certain outcome that will be automatically executed once certain parameters are fulfilled, players could agree to a smart contract that would deter them from defecting on the regulations, e.g., automatic release of national secrets upon defection on contract. By retroactively binding oneself via a smart contract that couples a deterring outcome with a certain action, one changes the payoff structure of players. Rather than having a potential positive payoff for defecting and developing AGI first, the payoff is turned negative by combining it with an undesired, automatically enforced penalty. By adding a negative outcome to defecting, perhaps AI safety could be turned into a problem resembling MAD (Mutually Assured Destruction), in which no player has an incentive to attack first due to the automatically enforced, collateral consequences that are tied to the move. Problems with this approach are that the automatic

execution of negative outcomes via a smart contract invites hacking, other abuse, or is vulnerable to flaws, which were risks that were discussed in relation to the Doomsday Device during the Cold war (Kahn, 1996).

■ Conceptualizing AI safety as common good: Apart from discouraging parties from defecting via punishing defectors, cooperation should be incentivized via the positive outcome that can be achieved as product of cooperation. Simply ideologically rebranding the AI-quest as a cooperative endeavor, rather than a race could have a positive psychological and motivational effect: Historical examples of increased cooperation via focus on the common good that could be created include the creation of CERN, ITER and the National Academies of Sciences. However, while avoiding adversarial language and focusing on the common good—a "CERN for AI"—could be helpful to achieve cooperation, it will not suffice given the strong incentives to cheat and move first and the sweeping effects of the Unilateralist's Curse (Bostrom, 2013).

Conclusion

The following summarizes the most important findings of the meeting and closes with recommendations regarding policy and how to move forward within the AI safety community.

Timelines

When considering AGI timelines, rather than prioritizing specific dates, what is more informative are the underlying reasons and models for arriving at a given date, including minority views if reasonable. Estimated timelines and recommendations varied widely, including both advocacy of a slower pace and a faster one. In the most extreme example of a slower pace, one could attempt to incentivize society to not make AGI at all by creating a framework of narrow AIs that is a good alternative and that cannot generalize. However, because players who are closest to AGI call the shots, persistent coordination problems among players are likely because narrow AI could at least be perceived to be the eternal second-best option.

In the case of a faster pace, one could attempt to speed up AGI development to bridge the acute risk period in AI, in which different players could engage in an arms-race toward AGI. Speeding up the inherent advantage that a small number of actors currently have to create a minimum viable AGI would avoid an increasingly larger pool of actors joining the race toward AGI. However, these unipolar scenarios might incentivize other players to threaten the first-mover, increasing the likelihood of other X-risks, e.g., nuclear war. Given that recommended timelines varied from "never" to "as soon as possible," more work is required to compare the advantages and disadvantages associated with different timelines.

Risk scenario I: Cybersecurity flaws

Cybersecurity is an integral part of AGI safety and should be recognized as such. While the EFF has initiated some efforts on cybersecurity, the field is neglected compared to the impact it can have: Without a secure architecture, an AGI cannot necessarily be expected to execute instructions reliably, because it is vulnerable to hacks, bugs, or could potentially take advantage of the vulnerability itself. Even without its relationship to AGI safety, cybersecurity is an important factor to consider, as most of today's infrastructure and economy are built on insecurable cyber

foundations. Insofar as we care about AGI risk due to its potential effects on human well-being, cybersecurity vulnerabilities that can be compromised to lead catastrophes, e.g., large-scale power outages with potentially catastrophic consequences, should be considered as well. Since cyber vulnerabilities already exist today, reducing cyber risk is urgent, especially since defense against cyber attacks could get disadvantaged compared to offense with increased use of AI techniques for cyber attacks. Potential solutions span defense-focused competitions for hackers, to deadline-backed government policies that incentivize companies to attend to their vulnerabilities, to learning from novel entities such as the blockchain ecosystem that naturally weed out insecure subsystems.

RIsk Scenario II: Near-term AI issues

On the path to AGI, narrow AI techniques will become ubiquitous and increasingly penetrate every aspect of human lives. Five areas were given special consideration: Autonomous Weapons; corporate concentration of power; phase change in science and engineering; surveillance, control, social engineering; and major unrest from economic disparity. While AI is a factor that could exacerbate the risks to human well-being inherent in all of the five areas, in some areas it could also be used for risk-reduction. For example in the case of surveillance, control, and social engineering, one prominent concern was that AI techniques could be used to create fraudulent data that is so realistic that they render it impossible to distinguish truths from untruths. Losing a common ground for truth would have great detrimental effects on all other areas considered as it makes communication and cooperation very difficult. Thus, we should focus on supporting AI techniques that are truth-revealing rather than truth-obscuring. Ultimately, AI-based Natural Language Processing (NLP) or sentiment analysis could highlight human emotional influence or biases in texts, even though current sentiment analysis tools run danger of being biased themselves if they are trained on biased data (Thompson, 2017).

Risk Scenario III: AGI concerns: Cooperation and Great Powers

Attention in AI safety concerns has shifted slightly from focusing on technical considerations of building safe AI systems to including strategic questions of cooperation among the different players involved in AI. One underlying rationale for this change is that some researchers assume, that much of the technical work required to solve AI safety can only be done closer to AGI take-off, because it is dependent on the type of system that is being built, even though there is disagreement on this assumption. At that future point, it would be very useful to have the necessary time available to do the required diligence on technical work. Thus, solving AI coordination problems that would allow the parties involved to slow development and focus on safety work, rather than cutting corners in an AI race, is an important focus now. While there is a lot of disagreement on whether a unipolar world or multipolar world of AI development and deployment is more secure, several coordination strategies for multipolar worlds are worthy of further investigation: Data-trade as naturally increasing cooperation, classifying permitted AI levels to agree on regulation, establishing a global council on regulation of AGI, smart contracts to enforce cooperation, and conceptualizing AI safety as common good. Relatedly, while there is some disagreement on how much and for how long organizations should open-source their research to be available to everyone, it would be helpful to create a more inclusive and cooperative publishing culture globally, e.g., especially for Chinese corporate research culture, which is still relatively closed compared to the US. Researchers could play a more active role in advocating a more collaborative approach and openness by making such an effort be a condition of employment when accepting new positions.

Conclusion

Government recommendations

Apart from the more specific government recommendations listed above in each scenario, more general recommendations include:

- Acknowledge and investigate the importance of near-term threats, especially for areas as cybersecurity, AI for Autonomous Weapons development, and AI for automation and the potential increase of technological unemployment.
- Foster a collaborative, open, positive culture among national organizations and international organizations, focused on the common good, rather than on adversarial aspects.
- Increase funding and support for AI safety organizations.

Moving forward

This meeting was one of only a few meetings of representatives of a diverse set of important AI safety organizations, others being CSRBAI, Beneficial AI, and Envisioning and Addressing Adversarial Outcomes. While some immediate recommendations were made, the meeting's main intention was not to reach definite consensus on specific topics but to open up communication on topics of high importance. The majority of participants were of the opinion that a more regular timing of such meetings would be useful. Of the future topics for meetings that were proposed by participants, those that were deemed especially useful are:

- Coordination, especially among great powers (including unipolar vs. multipolar scenarios, open vs. closed research, and collaboration of AI safety organizations)
- Cybersecurity as precondition for AI safety, and as immediate risk in itself
- Specific AGI take-off scenarios (early vs. delayed, seed AGI, agent-specific alignment problems)

Apart from working on the meeting-specific topics, future meetings would generally help to foster a more cooperative culture in the AI safety community. Exchanging strategies, state-of-the-art research, and synchronizing efforts would avoid having different organizations duplicate work, or work that could counter other organizations' work (a danger described in Bostrom's Unilateralist's Curse, [2013]) and would create an informed, and proactive research cohort (Conn, 2017). Increased collaboration among AI and AI safety researchers is in line with the Asilomar Principles and is one of the important factors that will prove beneficial across a variety of different AI risks scenarios.

References

Adler, S. (2017). "Breaking News". Radiolab. Available at: http://www.radiolab.org/story/breaking-news/

Armstrong, S., Sotala, K. (2012). "How We're Predicting AI - Or Failing To." MIRI. Available at: https://intelligence.org/files/PredictingAI.pdf

Baum, S. (2017). "A Survey of Artificial General Intelligence Projects for Ethics, Risk, Policy". Global Catastrophic Risk Institute Working Paper. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3070741

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

Grace, K. Salvatier, J. Dafoe, A. Zhang, B., Evans, O. (2017). "When Will AI Exceed Human Performance? Evidence from AI Experts." Computer Science. Available at: https://arxiv.org/pdf/1705.08807.pdf

Conn, A. (2017). "Safe Artificial Intelligence May Start with Collaboration". Future of Life Institute. Available at: https://futureoflife.org/2017/07/25/research-culture-principle/

Cotton-Barratt, O., Synder-Beattie, A. , Farquhar, S. (2016). "Beyond risk-benefit analysis: pricing externalities for gain-of-function research of concern." Future of Humanity Institute. Available at: https://www.fhi.ox.ac.uk/wp-content/uploads/GoFv9-1.pdf

Duettmann, A. (2017). Why Existential Risks Matter - And Some Decentralize Mitigation Strategies. Talk at BIL 2017. Available at: https://www.youtube.com/watch?v=SX79tofO5Bg

Eckersley, P. Nasser, Y. (2017). "AI Progress Measurement." Electronic Frontier Foundation. Available at: https://www.eff.org/ai/metrics

Gillula, J., Cardozo, N. Eckersley, P. (2016)." Does DARPA's Cyber Grand Challenge Need A Safety Protocol?" Electronic Frontier Foundation. Available at: https://www.eff.org/deeplinks/2016/08/darpa-cgc-safety-protocol

References

Kahle, B., (2014). "Corporations are 1st Generation AI's." Brewster Kahle Blog. Available at: http://brewster.kahle.org/2014/05/08/corporations-are-the-1st-generation-ais/

Davies, R. (2017). "Drones And The IED Threat." Action on Armed Violence. Available at: https://aoav.org.uk/2017/drones-ied-threat/

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

Drexler, E. (2017). "Comprehensive AI services neither require nor entail AGI." Available at: https://docs.google.com/document/d/12_Abdc_FFXh35n8QBg51NxgIN18qK-AAUT2RgBtPQLY/edit

Drexler, E. (2015). "MDL Intelligence Distillation: Exploring strategies for safe access to superintelligent problem-solving capabilities", Technical Report #2015-3, Future of Humanity Institute. Available at: https://www.fhi.ox.ac.uk/reports/2015-3.pdf

Goertzel, B. (2012). "The real reasons why we don't have AGI yet." Kurzweil.net. Available here: http://www.kurzweilai.net/the-real-reasons-we-dont-have-agi-yet

Bass, D. (2017). "AI Scientists gather to plot Doomsday Scenarios (And Solutions)." Bloomberg. https://www.bloomberg.com/news/articles/2017-03-02/ai-scientists-gather-to-plot-doomsday-scenarios-and-solutions

Bostrom, N. Sandberg, A., Douglas, T. (2013). "Unilateralist's Curse." Future of Humanity Institute. Available at: https://nickbostrom.com/papers/unilateralist.pdf

Bostrom, N., Dafoe, A., Flynn, C. (2016). "Policy Desiderata in the Development of Machine Superintelligence." Future of Humanity Institute. Available at:
https://nickbostrom.com/papers/aipolicy.pdf

Lobo, D., Levin, M. (2015). "Inferring Regulatory Networks from Experimental Morphological Phenotypes." PLOS Computational Biology. Available at: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004295#sec002

Kahn, Herman.. (1960). On Thermonuclear War. Princeton University Press.

Muehlhauser, L. (2016). "What do we know about AI timelines?" OpenPhilanthropy Project. Available at: http://www.openphilanthropy.org/focus/global-catastrophic-risks/potential-risks-advanced-artificial-intelligence/ai-timelines

Muehlhauser, L. (2013). "What is AGI?" MIRI. Available at: https://intelligence.org/2013/08/11/what-is-agi/

Peterson, C., Miller, M., Duettmann, A. (2017). "Cyber, Nano, and AGI Risks: Decentralized Approaches to Reducing Risks." Proceedings of First Colloquium on Catastrophic and Existential Risk. OnlineAvailable at: https://research.google.com/pubs/pub46290.html

References

Peterson, C. (2017). Cyber Risk: My Favorite Catastrophic Risk And How To Fix It". BIL talk, aAvailable at: https://www.youtube.com/watch?v=lEiUYseB9Z8

Rationalwiki. List of cognitive biases. https://rationalwiki.org/wiki/List_of_cognitive_biases

Rashid, F. (2015). "Cyber Attack on Power Grid Could Top $1 Trillion In Damage: Report." Security Week. Available at: http://www.securityweek.com/cyber-attack-power-grid-could-top-1-trillion-damage-report

Soares, N. (2017). "Ensuring Smarter-Than-Human Intelligence Has a Positive Outcome". MIRI. Available at: https://intelligence.org/2017/04/12/ensuring/

Tetlock P. Gardner, P. (2016). Superforecasting - The Art & Science of Prediction. Crown Publishers.

Thompson, A. (2017). "Google Is Sorry It's Sentiment Analyzer is Biased." Vice. Online at: https://motherboard.vice.com/en_us/article/ne3nkb/google-artificial-intelligence-bias-apology

Yampolskiy, R. (2016). "Artificial Intelligence Safety and Cyber Security: A Timeline of AI failure". Cornell University Library. Available at: https://arxiv.org/abs/1610.07997

Yudkowksy, E. (2017). "The AI Alignment Problem and Why It's Hard." MIRI. Available at: https://intelligence.org/2016/12/28/ai-alignment-why-its-hard-and-where-to-start/

FORESIGHT
INSTITUTE