Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

AGI Safety Literature Review

Tom Everitt, Gary Lea, Marcus Hutter
Australian National University
{tom.everitt, gary.lea, marcus.hutter}@anu.edu.au

Abstract

The development of Artificial General Intelligence (AGI) promises to be a major event. Along with its many potential benefits, it also raises serious safety concerns. The intention of this paper is to provide an easily accessible and up-to-date collection of references for the emerging field of AGI safety. A significant number of safety problems for AGI have been identified. We list these, and survey recent research on solving them. We also cover works on how best to think of AGI from the limited knowledge we have today, predictions for when AGI will first be created, and what will happen after its creation. Finally, we review the current public policy on AGI.

1 Introduction

An Artificial General Intelligence (AGI) is an AI system that equals or exceeds human intelligence in a wide variety of cognitive tasks. This is in contrast to today's AI systems that solve only narrow sets of tasks. Future AGIs may pose significant risks in addition to their many potential benefits [Bostrom, 2014]. The goal of this paper is to survey the literature relevant to these risks and their prevention.

Why study the safety of AGI before it exists, and before we even know whether it will ever exist? There are at least two types of reasons for this. The first is pragmatic. If AGI is created, and we do not know how to control it, then the outcome could be catastrophic [Bostrom, 2014]. It is customary to take precautions not only against catastrophes we know will happen, but also against catastrophes that have only a slight chance of occurring (for example, a city may decide to build earthquake safe buildings, even if the probability of an earthquake occurring is fairly low). As discussed in Section 3, AGI has more than a small probability of occurring, and it can cause significant catastrophes.

The second reason is scientific. Potential future AGIs are theoretically interesting objects, and the question of how humans can control machines more intelligent than themselves is philosophically stimulating. Section 2 summarizes progress made on understanding AGI, and Sections 4 and 5 consider ways in which this understanding has helped us to identify problems and generate solutions.

An extensive survey of the AGI safety literature was previously made by Sotala and Yampolskiy [2014]. Since then, the field has grown significantly. More up-to-date references are provided by this article, and by a number of recent research agendas and problem collections [Russell *et al.*, 2016; Amodei *et al.*, 2016; Leike *et al.*, 2017; Stoica *et al.*, 2017; Soares and Fallenstein, 2017; Taylor *et al.*, 2016]. A recent inventory of AGI projects and their attitudes towards ethics and safety also contributes to an overview of AGI safety research and attitudes [Baum, 2017].

This paper is structured as follows. Progress on how to think about yet-to-be-designed future AGI's is described in the first section (Section 2). Based partly on this understanding, we next survey predictions for when AGI will be created and what will happen after its creation (Section 3). We list and discuss identified AGI safety problems (Section 4), as well as proposals for solving or mitigating them (Section 5). Finally, we review the current public policy on AGI safety issues (Section 6), before making some concluding remarks (Section 7).

2 Understanding AGI

A major challenge for AGI safety research is to find the right conceptual models for plausible AGIs. This is especially challenging since we can only guess at the technology, algorithms, and structure that will be used. Indeed, even if we had the blueprint of an AGI system, understanding and predicting its behavior might still be hard: Both its design and its behavior could be highly complex. Nonetheless, several abstract observations and predictions are possible to make already at this stage.

2.1 Defining Intelligence

Legg and Hutter [2007] propose a formal definition of intelligence based on algorithmic information theory and the AIXI theory [Hutter, 2005]. They compare it to a large number of previously suggested definitions [Legg, 2007a]. Informally, their definition states that:

> "Intelligence measures an agent's ability to achieve goals in a wide range of environments."

The definition is non-anthropomorphic, meaning that it can be applied equally to humans and artificial agents. All present-day AIs are less intelligent than humans according to this def-

inition, as each AI is unable to achieve goals beyond a rather narrow domain. These domains can be for example ATARI environments [Mnih *et al.*, 2015, 2016; Hessel *et al.*, 2017], board-games [Silver *et al.*, 2016, 2017b,a], car-driving [Bojarski *et al.*, 2016; Huval *et al.*, 2015]. However, a trend towards greater generality can be observed, with e.g. car driving being a more general task than Chess, and AlphaZero simultaneously achieving state of the art performance on several challenging board games [Silver *et al.*, 2017a].

Following the Legg-Hutter definition, we may expect that a future, super-human AGI will be able to achieve more goals in a wider range of environments than humans. The most intelligent agent according to this definition is AIXI, which has been studied both mathematically and empirically; see Everitt and Hutter [2018b]; Leike [2016]; Hutter [2012b, 2005] for surveys. Safety work derived from AIXI is reviewed mostly in Section 5.

The Legg-Hutter intelligence definition measures what matters for control. The more intelligent an agent is, the more control it will have over aspects of the environment relating to its goals. If two agents with significantly different Legg-Hutter intelligence have conflicting goals in a shared environment, then the more intelligent of the two will typically succeed and the less intelligent fail. This points to the risks with increasingly intelligent AGIs: If their goals are not aligned with ours, then there will likely be a point where their goals will be achieved to the loss of ours [Russell, 2016].

2.2 Orthogonality

Bostrom's [2012; 2014] *orthogonality thesis* states that essentially any level of intelligence is compatible with any type of goal. Thus it does not follow, as is sometimes believed, that a highly intelligent AGI will realize that a simplistic goal such as creating paperclips or computing decimals of π is dumb, and that it should pursue something more worthwhile such as art or human happiness. Relatedly, Hume [1738] argued that *reason* is the slave of *passion*, and that a passion can never rationally be derived. In other words, an AGI will employ its intelligence to achieve its goals, rather than conclude that its goals are pointless. Further, if we want an AGI to pursue goals that we approve of, we better make sure that we design the AGI to pursue such goals: Beneficial goals will not emerge automatically as the system gets smarter.

2.3 Convergent Instrumental Goals

The orthogonality thesis holds for the *end goals* of the system. In stark contrast, the *instrumental goals* will often coincide for many agents and end goals [Omohundro, 2008; Bostrom, 2012]. Common instrumental goals include:

- Self-improvement: By improving itself, the agent becomes better at achieving its end goal.
- Goal-preservation and self-preservation: By ensuring that future versions of itself pursue the same goals, the end goal is more likely to be achieved.
- Resource acquisition: With more resources, the agent will be better at achieving the end goals.

Exceptions exist, especially in game-theoretic situations where the actions of other agents may depend on the agent's goals or other properties [LaVictoire *et al.*, 2014]. For example, an agent may want to change its goals so that it always chooses to honor contracts. This may make it easier for the agent to make deals with other agents.

2.4 Formalizing AGI

Bayesian, history-based agents have been used to formalize AGI in the so-called AIXI-framework [Hutter 2005; also discussed in Section 2.1]. Extensions of this framework have been developed for studying goal alignment [Everitt and Hutter, 2018a], multi-agent interaction [Leike *et al.*, 2016], space-time embeddedness [Orseau and Ring, 2012], self-modification [Orseau and Ring, 2011; Everitt *et al.*, 2016], observation modification [Ring and Orseau, 2011], self-duplication [Orseau, 2014a,b], knowledge seeking [Orseau, 2014], decision theory [Everitt et al., 2015], and death and suicide [Martin *et al.*, 2016].

Some aspects of reasoning are swept under the rug by AIXI and Bayesian optimality. Importantly, probability theory assumes that agents know all the logical consequences of their beliefs [Gaifman, 2014]. An impressive model of *logical non-omniscience* has recently been developed by Garrabrant [2016, 2017]. Notably, Garrabrant's theory avoids Gödelian obstacles for agents reasoning about improved versions of themselves [Fallenstein and Soares, 2014]. There is also hope that it can provide the foundation for a decision theory for logically uncertain events, such as how to bet on the 50th digit of π before calculating it.

2.5 Alternate Views

Drexler [private communication, 2017] argues that an AGI does not need to be an *agent* that plans to achieve a goal. An increasingly automatized AI research and development process where more and more of AI development is being performed by AI tools can become super-humanly intelligent without having any agent subcomponent. Avoiding to implement goal-driven agents that make long-term plans may avoid some safety concerns. Drexler [2015] outlines a theoretical idea for how to keep AIs specialized. Relatedly, Weinbaum and Veitas [2016], criticize the (rational) agent assumption underpinning most AGI theory.

However, Bostrom [2014, Ch. 10] and Gwern [2016], worry that the incentives for endowing a specialized *tool AI* with more general capabilities will be too strong. The more tasks that we outsource to the AI, the more it can help us. Thus, even if it were possible in theory to construct a safe tool AI, we may not be able to resist creating an agent AGI, especially if several competing organizations are developing AI and trying to reap its benefits. It is also possible that a system of tool AIs obtain agent properties, even if all of its subcomponents are specialized tool AIs.

3 Predicting AGI Development

Based on historical observations of economical and technological progress, and on the growing understanding of potential future AGIs described in Section 2, predictions have been made both for when the first AGI will be created, and what will happen once it has been created.

3.1 When Will AGI Arrive?

There is an ongoing and somewhat heated debate about when we can expect AGI to be created, and whether AGI is possible at all or will ever be created. For example, by extrapolating various technology trends until we can emulate a human brain, Kurzweil [2005] argues that AGI will be created around 2029. Chalmers [2010] makes a more careful philosophical analysis of the brain-emulation argument for AI, and shows that it defeats and/or avoids counter arguments made by Lucas [1961], Dreyfus [1972], Searle [1980], and Penrose [1994]. Chalmers is less optimistic about the timing of AGI, and only predicts that it will happen within this century.

Surveys among AI researchers have found median predictions for AGI between 2040 and 2061, with estimates varying widely, from never to just a few years into the future [Baum *et al.*, 2011; Müller and Bostrom, 2016; Grace *et al.*, 2017]. Algorithmic progress have been tracked by Grace [2013], Eck ersley and Nasser [2018], and AI Impacts [2018b], and the costs of computing have been tracked by AI Impacts [2018b]. Notably, the computing power available for AI has doubled roughly every 3-4 months in recent years [Amodei and Hernandez, 2018]. A new MIT course on AGI shows that the AGI prospect is becoming more mainstream [Fridman, 2018]. Stanford has a course on AI safety [Sadigh, 2017]. Jilk [2017] argues that an AGI must have a conceptual-linguistic faculty in order to be able to access human knowledge or interact effectively with the world, and that the development of systems with conceptual-linguistic ability can be used as an indicator of AGI being near.

3.2 Will AGI Lead to a Technological Singularity?

As explained in Section 2.3, one of the instrumental goals of almost any AGI will be self-improvement. The greater the improvement, the likelier the end goals will be achieved. This can lead to *recursive* self-improvement, where a self-upgraded AGI is better able to find yet additional upgrades, and so on. If the pace of this process increases, we may see an *intelligence explosion* once a critical level of self-improvement capability has been reached [Good, 1966; Vinge, 1993; Kurzweil, 2005; Yudkowsky, 2008; Hutter, 2012a; Bostrom, 2014]. Already John von Neumann have been quoted calling this intelligence explosion a *singularity* [Ulam, 1958]. Singularity should here not be understood in its strict mathematical sense, but more loosely as a point where our models break.

Some counter arguments to the singularity have been structured by Walsh [2016], who argues that an intelligence explosion is far from inevitable:

- Intelligence measurement: The singularity predicts an increasingly rapid development of intelligence. However, it is not quite clear how we should measure intelligence [Hutter, 2012]. A rate of growth that looks fast or exponential according to one type of measurement, may look ordinary or linear according to another measurement (say, the log-scale).
- Fast thinking dog: No matter how much we increase the speed at which a dog thinks, the dog will never beat a decent human at chess. Thus, even if computers keep getting faster, this alone does not entail their ever becoming smarter than humans.
- Anthropocentric: Proponents of the singularity often believes that somewhere around the human level of intelligence is a critical threshold, after which we may see quick recursive self-improvement. Why should the human level be special?
- Meta-intelligence, diminishing returns, limits of intelligence, computational complexity: It may be hard to do self-improvement or be much smarter than humans due to a variety of reasons, such as a fundamental (physical) upper bound on intelligence or difficulty of developing machine learning algorithms.

These arguments are far from conclusive, however. In *Life 3.0*, Tegmark [2017] argues that AGI constitutes a third stage of life. In the first stage, both hardware and software is evolved (e.g. in bacteria). In the second stage, the hardware is evolved but the software is designed. The prime example is a human child who goes to school and improves her knowledge and mental algorithms (i.e. her software). In the third stage of life, both the software and hardware is designed, as in an AGI. This may give unprecedented opportunities for quick development, countering the anthropocentric argument by Walsh. In relation to the limits of intelligence arguments, Bostrom [2014] argues that an AGI may think up to a million times faster than a human. This would allow it to do more than a millennium of mental work in a day. Such a speed difference would make it very hard for humans to control the AGI. Powerful mental representations may also allow an AGI to quickly supersede human intelligence in quality Sotala [2017], countering the fast-thinking dog argument. The possibility of brain-emulation further undermines the fast-thinking dog argument. Yampolskyi [2017] also replies to Walsh's arguments.

Kurzweil's [2005] empirical case for the singularity has been criticized for lack of scientific rigor [Modis, 2006]. Modis [2002] argues that a logistic function fits the data better than an exponential function, and that logistic extrapolation yields that the rate of complexity growth in the universe should have peaked around 1990.

In conclusion, there is little consensus on whether and when AGI will be created, and what will happen after its creation. Anything else would be highly surprising, given that no similar event have previously occurred. Nonetheless, AGI being created within the next few decades and quickly superseding human intelligence seems like a distinct possibility.

4 Problems with AGI

Several authors and organizations have published research agendas that identify potential problems with AGI. Russell *et al.* [2016] and the Future of Life Institute (FLI) take the broadest view, covering societal and technical challenges in both the near and the long term future. Soares and Fallenstein [2017] at the Machine Intelligence Research Institute (MIRI) focus on the mathematical foundations for AGI, including decision theory and logical non-omniscience. Several subsequent agendas and problem collections try to bring the

sometimes "lofty" AGI problems down to concrete machine learning problems: Amodei *et al.* [2016] at OpenAI et al., Leike *et al.* [2017] at DeepMind, and Taylor *et al.* [2016] also at MIRI. In the agenda by Stoica *et al.* [2017] at UC Berkeley, the connection to AGI has all but vanished. For brevity, we will refer to the agendas by the organization of the first author, with MIRI-AF the agent foundations agenda by Soares and Fallenstein [2017] and MIRI-ML the machine learning agenda by [Taylor *et al.*, 2016]. Figure 1 shows some connections between the agendas. Figure 1 also makes connections to research done by other prominent AGI safety institutions: Oxford Future of Humanity Institute (FHI), Australian National University (ANU), and Center for Human-Compatible AI (CHAI).

Some clusters of problems appear in multiple research agendas:

- Value specification: How do we get an AGI to work towards the right goals? MIRI calls this value specification. Bostrom [2014] discusses this problem at length, arguing that it is much harder than one might naively think. Davis [2015] criticizes Bostrom's argument, and Bensinger [2015] defends Bostrom against Davis' criticism. Reward corruption, reward gaming, and negative side effects are subproblems of value specification highlighted in the DeepMind and OpenAI agendas.
- Reliability: How can we make an agent that keeps pursuing the goals we have designed it with? This is called *highly reliable agent design* by MIRI, involving decision theory and logical omniscience. DeepMind considers this the self-modification subproblem.
- Corrigibility: If we get something wrong in the design or construction of an agent, will the agent cooperate in us trying to fix it? This is called error-tolerant design by MIRI-AF and *corrigibility* by Soares *et al.* [2015]. The problem is connected to safe interruptibility as considered by DeepMind.
- Security: How to design AGIs that are robust to adversaries and adversarial environments? This involves building sandboxed AGI protected from adversaries (Berkeley), and agents that are robust to adversarial inputs (Berkeley, DeepMind).
- Safe learning: AGIs should avoid making fatal mistakes during the learning phase. Subproblems include safe exploration and distributional shift (DeepMind, OpenAI), and continual learning (Berkeley).
- Intelligibility: How can we build agent's whose decisions we can understand? Connects explainable decisions (Berkeley) and informed oversight (MIRI). DeepMind is also working on these issues, see Section 5.5 below.
- Societal consequences: AGI will have substantial legal, economic, political, and military consequences. Only the FLI agenda is broad enough to cover these issues, though many of the mentioned organizations evidently care about the issue [Brundage *et al.*, 2018; DeepMind, 2017].

There are also a range of less obvious problems, which have received comparatively less attention:

- Subagents: An AGI may decide to create *subagents* to help it with its task [Soares *et al.* 2015, Orseau, 2014a,b]. These agents may for example be copies of the original agent's source code running on additional machines. Subagents constitute a safety concern, because even if the original agent is successfully shut down, these subagents may not get the message. If the subagents in turn create subsubagents, they may spread like a viral disease.
- Malign belief distributions: Christiano [2016] argues that the *universal distribution* M [Solomonoff, 1964a,b, 1978; Hutter, 2005] is malign. The argument is somewhat intricate, and is based on the idea that a hypothesis about the world often includes simulations of other agents, and that these agents may have an incentive to influence anyone making decisions based on the distribution. While it is unclear to what extent this type of problem would affect any practical agent, it bears some semblance to aggressive *memes*, which do cause problems for human reasoning [Dennet, 1990].
- Physicalistic decision making: The *rational agent* framework is pervasive in the study of artificial intelligence. It typically assumes that a well-delineated entity interacts with an environment through action and observation channels. This is not a realistic assumption for *physicalistic* agents such as robots that are part of the world they interact with [Soares and Fallenstein, 2017].
- Multi-agent systems: An artificial intelligence may be copied and distributed, allowing instances of it to interact with the world in parallel. This can significantly boost learning, but undermines the concept of a single agent interacting with the world.

While these problems may seem esoteric, a security mindset [Yudkowsky, 2017] dictates that we should not only protect ourselves from things that can clearly go wrong, but also against anything that is not guaranteed to go right. Indeed, unforeseen errors often cause the biggest risks. For this reason, the biggest safety problem may be one that we have not thought of yet – not because it would necessarily be hard to solve, but because in our ignorance we fail to adopt measures to mitigate the problem.

5 Design Ideas for Safe AGI

We next look at some ideas for creating safe AGI. There is not always a clear line distinguishing ideas for safe AGI from other AI developments. Many works contribute to both simultaneously.

5.1 Value Specification

RL and misalignment. Reinforcement learning (RL) [Sutton and Barto, 1998] is currently the most promising framework for developing intelligent agents and AGI. Combined with Deep Learning, it has seen some remarkable recent successes, especially in playing board games [Silver *et al.*, 2016,

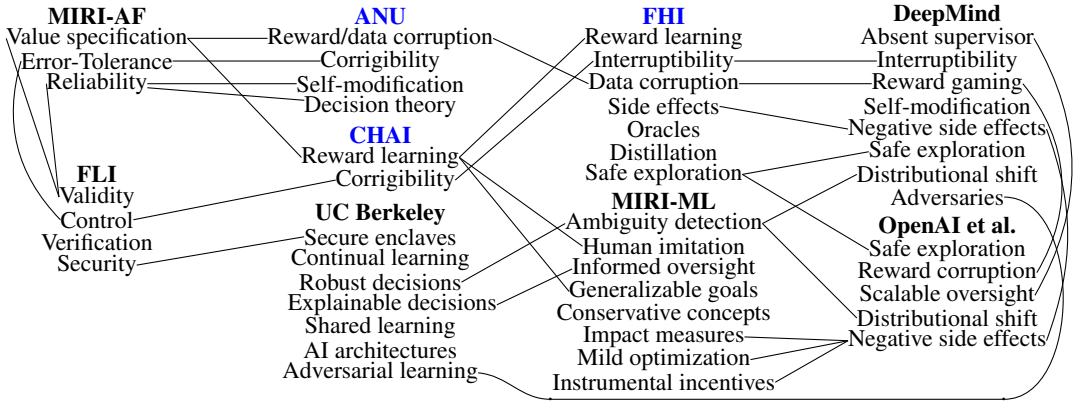

Figure 1: Connections between problems stated in different AGI safety research agendas (for ANU, CHAI, and FHI, the agendas are inferred from their recent publications).

2017a,b] and computer games [Mnih *et al.*, 2015, 2016; Hessel *et al.*, 2017].

Aligning the goals of an RL agent with the goals of its human supervisor comprises significant challenges, however [Everitt and Hutter, 2018]. These challenges include correct specification of the reward function, and avoiding that the agent takes shortcuts in optimizing it. Such shortcuts include the agent corrupting the observations on which the reward function evaluates performance, modifying the reward function to give more reward, hijacking the reward signal or the memory location of the reward, and, in the case of an interactively learned reward function, corrupting the data training the reward function. Everitt and Hutter [2018] categorize misalignment problems in RL, and suggest a number of techniques for managing the various sources of misalignment. The rest of this subsection reviews other work that has been done on designing agents with correctly specified values.

Learning a reward function from actions and preferences. One of the main challenges in scaling RL to the real world includes designing the reward function. This is particularly critical for AGI, as a poorly designed reward function would likely lead to a misaligned agent. As an example of misalignment, Clark and Amodei [2016] found that their boat racing agent preferred going in circles and crashing into obstacles instead of winning the race, due to a subtly misspecified reward function. Lehman *et al.* [2018], Gwern [2011], and Irpan [2018] have many more examples. Analogous failures in AGIs could cause severe catastrophes. The DeepMind problem collection calls this a *reward gaming* problem. One potential way around the problem of gameable reward functions is to let the agent learn the reward function. This lets designers offload some of the design work to powerful machine learning techniques.

Inverse reinforcement learning (IRL) [Ng and Russell, 2000; Ziebart et al., 2008; Choi and Kim, 2011] is a framework for learning a reward function from the actions of an expert, often a human demonstrator. In one famous example, Abbeel *et al.* [2007] taught an agent acrobatic helicopter flight by observing the actions of a human pilot. Impressively, the agent ultimately became better at flying than the pilot it observed. However, a learned reward function cannot be better than the data that trained it. If all training happens before the agent is launched into the environment, then the data may not properly describe situations that the agent reaches far into its lifetime (a so-called *distributional shift* problem; Amodei *et al.* 2016). For this reason, interactive training of the reward function may be preferable, as it allows the training data to adapt to any new situation the agent may encounter.

Cooperative inverse reinforcement learning (CIRL) is a generalization of IRL that lets the expert and the agent act simultaneously in the same environment, with the agent interactively learning the expert's preferences [Hadfield-Menell *et al.*, 2016]. Among other things, this allows the expert to take demonstrative actions that are suboptimal according to his or her reward function but more informative to the agent, without the agent being led to infer an incorrect reward function. The CIRL framework can be used to build agents that avoid interpreting reward functions overly literally, thus avoiding some misalignment problems with RL [Hadfield-Menell *et al.*, 2017b].

A reward functions can also be learned from a human rating short video clips of (partial) agent trajectories against each other [Christiano *et al.*, 2017]. For example, if the human consistently rates scenarios where the agent falls off a cliff lower than other scenarios, then the learned reward function will assign a low reward to falling off a cliff. Using this technique, a non-expert human can teach an agent complex behaviors that would have been difficult to directly program a reward function for. Warnell [2017] use a related approach, needing only 15 minutes of human feedback to teach the agent the ATARI game Bowling. In order to scale this method to more complex tasks where evaluation is non-trivial, Irving *et al.* [2018] propose letting two systems debate which option is better, highlighting flaws in each others suggestions and arguments. Ideally, following the debate will significantly boost the human's ability to make an informed evaluation.

On a fundamental level, learning from actions and learning from preferences is not widely different. Roughly, a choice of action a over action b can be interpreted as a preference

for the future trajectories resulting from action a over the trajectories resulting from action b. However, a few notable differences can still be observed. First, at least in Christiano's [2017] framework, preferences always apply to *past* events. In contrast, an action in the CIRL framework typically gives information about which *future* events the human prefers. A drawback is that in order for the action to carry information about future events, the action must be chosen (somewhat) rationally. Humans do not always act rationally; indeed, we exhibit several systematic biases [Kahneman, 2011]. A naive application of (C)IRL therefore runs the risk of inferring an incorrect reward function. To address this, Evans *et al.* [2016] develop a method for learning the reward function of agents exhibiting some human-like irrationalities. Without assumptions on the type of irrationality the expert exhibits, nothing can be learned about the reward function [Armstrong and Mindermann, 2017]. In comparison, learning from preferences seems to require weaker rationality assumptions on the human's part, as correctly stating ones preferences may be easier than acting rationally.

Yet another approach to learning a reward function is to learn it from stories [Riedl and Harrison, 2016].

Approval-directed agents. In a series of blog posts, Christiano [2014] suggests that AGIs should be designed to maximize approval for their actions rather than trying to reach some goal. He argues that approval-directed systems have many of the same benefits of goal-directed systems while avoiding some of their worst pitfalls. Christiano [2015] and Cotra [2018] outline a method for how approval-directed agents can be chained together in a hierarchy, boosting the accuracy of the approvals of the human at the top of the chain.

Reward corruption. Reinforcement learning AGIs may hijack their reward signal and feed themselves maximal reward [Ring and Orseau, 2011]. Interestingly, model-based agents with preprogrammed reward functions are much less prone to this behavior [Everitt *et al.*, 2016; Hibbard, 2012]. However, if the reward function is learned online as discussed above, it opens up the possibility of *reward learning corruption*. An AGI may be tempted to influence the data training its reward function so it points towards simple-to-optimize reward functions rather than harder ones [Armstrong, 2015]. Everitt *et al.* [2017] show that the type of data the agent receives matter for reward learning corruption. In particular, if the reward data can be cross-checked between multiple sources, then the reward corruption incentive diminishes drastically. Everitt *et al.* also evaluate a few different approaches to reward learning, finding that the *human action*-data provided in CIRL is much safer than the *reward*-data provided in standard RL, but that CIRL is not without worrying failure modes.

Side effects. An AGI that becomes overly good at optimizing a goal or reward function that does not fully capture all human values, may cause significant negative side effects [Yudkowsky, 2009]. The *paperclip maximizer* that turns the earth and all humans into paperclips is an often used example [Bostrom, 2014], now available as a highly addictive computer game [Lantz, 2017]. Less extreme examples include companies that optimize profits and cause pollution and other externalities as negative side effects.

The most serious side effects seem to occur when a target function is optimized in the extreme (such as turning the earth into paperclips). Quantilization can avoid over-optimization under some assumptions [Everitt *et al.* 2017; Taylor, 2016]. Another more specific method is to "regularize" reward by the impact the policy is causing [Armstrong and Levinstein, 2017]. How to measure impact remains a major open question, however.

Connections to economics. The goal alignment problem has several connections to the economics literature. It may be seen as an instance of Goodhart's law [Goodheart, 1975] which roughly states that any measure of performance ceases to be a good measure once it is optimized for. Mannheim and Garrabrant [2018] categorize instances of Goodhart's law. It may also be seen as a principal-agent problem: The connections have been fleshed out by Hadfield-Menell and Hadfield [2018].

5.2 Reliability

Self-modification. Even if the reward function is correctly specified, an AGI may still be able to corrupt either the reward function itself or the data feeding it. This can happen either intentionally if such changes can give the agent more reward, or accidentally as a side effect of the agent trying to improve itself (Section 2.3).

A utility self-preservation argument going back to at least Schmidhuber [2007] and Omohundro [2008] says that agents should not want to change their utility functions, as that will reduce the utility generated by their future selves, as measured by the current utility function. Everitt *et al.* [2016] formalize this argument, showing that it holds under three nontrivial assumptions: (1) The agent needs to be model-based, and evaluate future scenarios according to its current utility function; (2) the agent needs to be able to predict how self-modifications affect its future policy; and (3) the reward function itself must not endorse self-modifications. In RL [Sutton and Barto, 1998], model-free agents violate the first assumption, off-policy agents such as Q-learning violate the second, and the third assumption may fail especially in learned reward/utility functions (Section 5.1). Hibbard [2012] and Orseau [2011] Hibbard [2012] and Orseau and Ring [2011] also study the utility self-preservation argument.

5.3 Corrigibility

By default, agents may resist shutdown and modifications due to the self-preservation drives discussed in Section 2.3. Three rather different approaches have been developed to counter the self-preservation drives.

Indifference. By adding a term or otherwise modifying the reward function, the agent can be made indifferent between some choices of future events, for example shutdown or software corrections [Armstrong, 2017]. For example, this tech-

nique can be used to construct variants of popular RL algorithms that do not learn to prevent interruptions [Orseau and Armstrong, 2016].

Ignorance. Another option is to construct agents that behave as if a certain event (such as shutdown or software modification) was certain not to happen [Everitt *et al.*, 2016]. For example, off-policy agents such as Q-learning behave as if they will always act optimally in the future, thereby effectively disregard the possibility that their software or policy be changed in the future. Armstrong [2017] show that ignorance is equivalent to indifference in a certain sense.

Uncertainty. In the CIRL framework [Hadfield-Menell *et al.*, 2016], agents are uncertain about their reward function, and learn about the reward function through interaction with a human expert. Under some assumptions on the human's rationality and the agent's level of uncertainty, this leads to naturally corrigible agents [Hadfield-Menell *et al.* 2017a; Wängberg, 2017]. Essentially, the agent will interpret the human's act of shutting them down as evidence that being turned off has higher reward than remaining turned on. In some cases where the human is likely to make suboptimal choices, the agent may decide to ignore a shut down command. There has been some debate about whether this is a feature [Milli *et al.*, 2017] or a bug [Carey, 2018].

Continuous testing. Arnold and Scheutz [2018] argue that an essential component of corrigibility is to detect misbehavior as early as possible. Otherwise, significant harm may be caused without available corrigibility equipment having been put to use. They propose an ethical testing framework that continually monitors the agent's behavior on simulated ethics tests.

5.4 Security

Adversarial counterexamples. Deep Learning [e.g. Goodfellow *et al.*, 2016] is a highly versatile tool for machine learning, and a likely building block for future AGIs. Unfortunately, it has been observed that small perturbations of inputs can cause severe misclassification errors [Szegedy et al., 2013; Goodfellow et al., 2014; Evtimov et al., 2017; Athalye et al., 2017].

In a recent breakthrough, Katz *et al.* [2017] extend the Simplex algorithm to neural networks with rectified linear units (Relus). Katz *et al.* call the extended algorithm ReluPlex, and use it to successfully verify the behavior of neural networks with 300 Relu nodes in 8 layers. They gain insight into the networks' behaviors in certain important regions, as well as the sensitivity to adversarial perturbations.

5.5 Intelligibility

While it is infamously hard to understand exactly what a deep neural network has learned, recently some progress has been made. DeepMind's *Psychlab* uses tests from psychology implemented in a 3D environment to understand deep RL agents. The tests led them to a simple improvement of the UNREAL agent [Leibo *et al.*, 2018] Zahavy *et al.* [2016] instead use the dimensionality reduction technique t-SNE on the activations of the top neural network layer in DQN [Mnih *et al.*, 2015]. revealing how DQN represents policies in ATARI games. Looking beyond RL, Olah *et al.* [2017] summarize work on visualization of features in image classification networks in a beautiful Distill post. Another line of work tries to explain what text and speech networks have learned Alvarez-Melis and Jaakkola, 2017; Belinkov and Glass, 2017; Lei et al., 2016].

5.6 Safe learning

During training, a standard Deep RL agent such as DQN commits on the order of a million catastrophic mistakes such as jumping off a cliff and dying [Saunders *et al.*, 2017]. Such mistakes could be very expensive if they happened in the real world. Further, we do not want an AGI to accidentally set off all nuclear weapons in a burst of curiosity late in its training phase. Saunders *et al.* [2017] propose to fix this by training a neural network to detect potentially catastrophic actions from training examples provided by a human. The catastrophe detector can then override the agent's actions whenever it judges an action to be too dangerous. Using this technique, they manage to avoid all catastrophes in simple settings, and a significant fraction in more complex environments. A similar idea was explored by Lipton *et al.* [2016]. Instead of using human-generated labels, their catastrophe detector was trained automatically on the agent's catastrophes. Unsurprisingly, this reduces but does not eliminate catastrophic mistakes. A survey over previous work on safe exploration in RL is provided by [Garcia and Fernandez, 2015].

6 Public Policy on AGI

Recommendations. In a collaboration spanning 14 organizations, Brundage *et al.* [2018] consider scenarios for how AI and AGI may be misused and give advice for both policy makers and researchers. Regulation of AI remains a controversial topic, however. On the one hand, Erdeleyi and Goldsmith [2018] call for global regulatory body. Others worry that regulations may limit the positive gains from AI [Gurkaynak *et al.* 2016; Nota, 2015], and recommend increased public funding for safety research [Nota, 2015]. Baum [2017b] is also wary of regulation, but for slightly different reasons. He argues that *extrinsic* measures such as regulations run the risk of backfiring, making AI researchers look for ways around the regulations. He argues that extrinsic measures such as regulations may backfire, and instead recommends working on social norms and other measures that make developers want to develop safe AI. Armstrong *et al.* [2016] counterintuitively find that information sharing between teams developing AGI exacerbates the risk of an AGI race.

Policy makers. Although public policy making is often viewed as the domain of public bodies, it should be remembered that many organizations such as corporations, universities and NGOs frequently become involved through advocacy, consulting, and joint projects. Indeed, such involvement can often extend to de facto or "private" regulation via

organizational guidelines, organizational policies, technical standards and similar instruments.

Professional organizations have already taken a leading role. The IEEE, for example, is developing guidelines on Ethically Aligned Design [IEEE, 2017a,b]. Meanwhile, the ACM and the SIGAI group of AAAI have co-operated to establish a new joint conference on AI, ethics and society, AIES. Economic policy and technical standards organizations have also started to engage: for example, the OECD has established a conference on "smart policy making" around AI developments [OECD, 2017] and ISO/IEC has established a technical committee on AI standards [ISO/IEC, 2017]. Corporations and corporate consortia are also involved, typically through the public-facing aspects of their own corporate policies [Intel, 2017; IBM, 2018] or through joint development of safety policies and recommendations which consortia members will adopt [Partnership on AI, 2016].

Finally, in addition to the traditional public roles of academia and academics, there are an increasing number of academically affiliated or staffed AI organizations. With varying degrees of specificity, these work on technical, economic, social and philosophical aspects of AI and AGI. Organizations include the Future of Humanity Institute (FHI), the Machine Intelligence Research Institute (MIRI), the Centre for the Study of Existential Risk (CSER) and the Future of Life Institute (FLI).

Current policy anatomy. It could be said that public policy on AGI does not exist. More specifically, although work such as Baum [2017] highlights the extent to which AGI is a distinct endeavor with its own identifiable risk, safety, ethics (RISE) issues, public policy AGI is currently seldom separable from default public policy on AI taken as a whole (PPAI). Existing PPAI are typically structured around (a) significant financial incentives (e.g. grants, public-private co-funding initiatives, tax concessions) and (b) preliminary coverage of ethical, legal and social issues (ELSI) with a view to more detailed policy and legislative proposals later on [Miller *et al.*, 2018; FTI Consulting, 2018].

In the case of the EU, for example, in addition to experimental regulation with its new algorithmic decision-making transparency requirements in [EUR-lex, 2016, Article 22, General Data Protection Regulation] its various bodies and their industry partners have committed over 3 billion Euro to AI and robotics R&D and engaged in two rounds of public consultation on the European Parliament's proposed civil law liability framework for AI and robotics [Ansip, 2018]. However, the much demanded first draft of an overarching policy framework is still missing, being slated for delivery by the European Commission no earlier than April 2018.

Elsewhere, spurred into action by the implications of the AlphaGo victory and China's recent activities (outlined below), South Korea and Japan have already rapidly commenced significant public and public-private investment programs together with closer co-ordination of state bodies, industry and academia [Ha, 2016; Volodzsko, 2017]. Japan is also additionally allowing experimental regulation in some economic sectors [Takenaka, 2017]. The UK has started work on a preliminary national policy framework on robotics and AI [UK Parliament, 2017; Hall and Presenti, 2017], and have established a national Centre for Data Ethics and Innovation [CSER, 2017].

Current policy dynamics. Although there is substantial positive co-operation between universities, corporations and other organizations, there is a negative dynamic operating the nation-state, regional and international context. Contrary to the expert recommendations above, there is increasing rhetoric around an AI "arms race" [Cave and OhEigeartaigh, 2018], typified by President Vladimir Putin's September 2017 comment that "... whoever becomes the leader in [the AI] sphere will become the leader in the world" [Apps, 2017]. Relatedly, China's 8 July 2017 AI policy announcement included being the global leader in AI technology by 2030 [PRC State Council, 2017; Kania, 2018; Ding, 2018]. It also included aims of "creating a safer, more comfortable and convenient society." Alongside this policy shift has been increased Sino-American competition for AI talent [Cyranoski, 2018]. In the US, the Obama Administration began consultation and other moves towards a federal policy framework for AI technology investment, development and implementation [Agrawal *et al.* 2016; White House OSTP, 2016]. However, the Trump Administration abandoned the effort to focus mainly on military spending on AI and cyber-security [Metz, 2018].

Policy outlook. Given the above, looking forward it would appear that the organizations noted above will have to work hard to moderate the negative dynamic currently operating at the nation-state, regional and international level. Useful guidance for researchers and others engaging with public policy and regulatory questions on AI is given by 80 000 Hours [2017]. Further references on public policy on AGI can be found in [Sotala and Yampolskiy 2014; Dafoe, 2017].

7 Conclusions

AGI promises to be a major event for humankind. Recent research has made important progress on how to think about potential future AGIs, which enables us to anticipate and (hopefully) mitigate problems before they occur. This may be crucial, especially if the creation of a first AGI leads to an "intelligence explosion". Solutions to safety issues often have more near-term benefits as well, which further adds to the value of AGI safety research.

It is our hope that this summary will help new researchers enter the field of AGI safety, and provide traditional AI researchers with an overview of challenges and design ideas considered by the AGI safety community.

References[1]

[Amodei *et al.*, 2016] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety, 2016.

[1]The full list of references is in the extended version `https://arxiv.org/abs/1805.01109`

[Armstrong *et al.*, 2016] Stuart Armstrong, Nick Bostrom, and Carl Shulman. Racing to the Precipice: A Model of Artificial Intelligence Development. *AI and Society*, 31(2):201–206, 2016.

[Baum, 2017] Seth D. Baum. A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy. Technical Report November, Global Catastrophic Risk Institute, 2017.

[Bostrom, 2012] Nick Bostrom. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, pages 1–16, 2012.

[Bostrom, 2014] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[Brundage *et al.*, 2018] Miles Brundage, Shahar Avin, *et al.* The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation, 2018.

[Chalmers, 2010] David J Chalmers. The singularity: A philosophical analysis. *Journal of Consciousness Studies*, 17(9-1):7–65, 2010.

[Christiano *et al.*, 2017] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *NIPS.*, pages 4302–4310, 2017.

[Davis, 2015] Ernest Davis. Ethical guidelines for a superintelligence. *Artificial Intelligence*, 220:121–124, 2015.

[Everitt *et al.*, 2016] Tom Everitt, Daniel Filan, Mayank Daswani, and Marcus Hutter. Self-modification of policy and utility function in rational agents. In *AGI*, pages 1–11, 2016.

[Everitt *et al.*, 2017] Tom Everitt, Victoria Krakovna, Laurent Orseau, Marcus Hutter, and Shane Legg. Reinforcement Learning with Corrupted Reward Signal. In *IJCAI*, pages 4705–4713, 2017.

[Garrabrant *et al.*, 2017] Scott Garrabrant, Tsvi Benson-Tilsen, Andrew Critch, Nate Soares, and Jessica Taylor. A Formal Approach to the Problem of Logical Non-Omniscience. *Electronic Proceedings in Theoretical Computer Science*, 251:221–235, 2017.

[Hadfield-Menell *et al.*, 2016] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart J Russell. Cooperative Inverse Reinforcement Learning. In *NIPS*, pages 3909–3917, 2016.

[Hadfield-Menell *et al.*, 2017a] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart J Russell. The Off-Switch Game. In *IJCAI*, pages 220–227, 2017.

[Hadfield-Menell *et al.*, 2017b] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. Inverse Reward Design. In *NIPS*, pages 6768–6777, 2017.

[Hutter, 2005] Marcus Hutter. *Universal Artificial Intelligence*. Springer-Verlag, Berlin, 2005.

[Hutter, 2012] Marcus Hutter. Can Intelligence Explode? *Journal of Consciousness Studies*, 19(1-2):143–146, 2012.

[Katz *et al.*, 2017] Guy Katz, Clark Barrett, David L. Dill, Kyle Julian, and Mykel J. Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. In *(CAV)*, pages 97–117, 2017.

[Kurzweil, 2005] Ray Kurzweil. *The Singularity Is Near*. Viking, 2005.

[Legg and Hutter, 2007] Shane Legg and Marcus Hutter. Universal Intelligence: A definition of machine intelligence. *Minds & Machines*, 17(4):391–444, 2007.

[Leike *et al.*, 2017] Jan Leike, Miljan Martic, *et al.* AI Safety Gridworlds, 2017.

[Milli *et al.*, 2017] Smitha Milli, Dylan Hadfield-Menell, Anca Dragan, and Stuart J Russell. Should robots be obedient? In *IJCAI*, pages 4754–4760, 2017.

[Omohundro, 2008] Stephen M Omohundro. The Basic AI Drives. In P. Wang, B. Goertzel, and S. Franklin, editors, *Artificial General Intelligence*, volume 171, pages 483–493. IOS Press, 2008.

[Orseau and Armstrong, 2016] Laurent Orseau and Stuart Armstrong. Safely interruptible agents. In *32nd Conference on Uncertainty in Artificial Intelligence.*, 2016.

[Orseau, 2014] Laurent Orseau. Universal Knowledge-seeking Agents. *Theoretical Computer Science*, 519:127–139, 2014.

[Ring and Orseau, 2011] Mark Ring and Laurent Orseau. Delusion, Survival, and Intelligent Agents. In *AGI*, pages 1–11. Springer, 2011.

[Russell *et al.*, 2016] Stuart J Russell, Daniel Dewey, and Max Tegmark. Research Priorities for Robust and Beneficial Artificial Intelligence. *AI Magazine*, 36(4), 2016.

[Soares and Fallenstein, 2017] Nate Soares and Benya Fallenstein. Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda. In V Callaghan *et al.* editors, *The Technological Singularity: Managing the Journey*, chapter 5, pages 103–125. Springer, 2017.

[Soares *et al.*, 2015] Nate Soares, Benya Fallenstein, Eliezer S Yudkowsky, and Stuart Armstrong. Corrigibility. In *AAAI Workshop on AI and Ethics*, pages 74–82, 2015.

[Sotala and Yampolskiy, 2014] Kaj Sotala and Roman V Yampolskiy. Responses to catastrophic AGI risk: a survey. *Physica Scripta*, 90(1), 2014.

[Sotala, 2017] Kaj Sotala. How feasible is the rapid development of artificial superintelligence? *Physica Scripta*, 92(11):1–29, 2017.

[Stoica *et al.*, 2017] Ion Stoica, Dawn Song, *et al.* A Berkeley View of Systems Challenges for AI. Technical report, EECS Department, University of California, Berkeley, 2017.

[Taylor *et al.*, 2016] Jessica Taylor, Eliezer S Yudkowsky, Patrick LaVictoire, and Andrew Critch. Alignment for Advanced Machine Learning Systems. Technical report, MIRI, 2016.