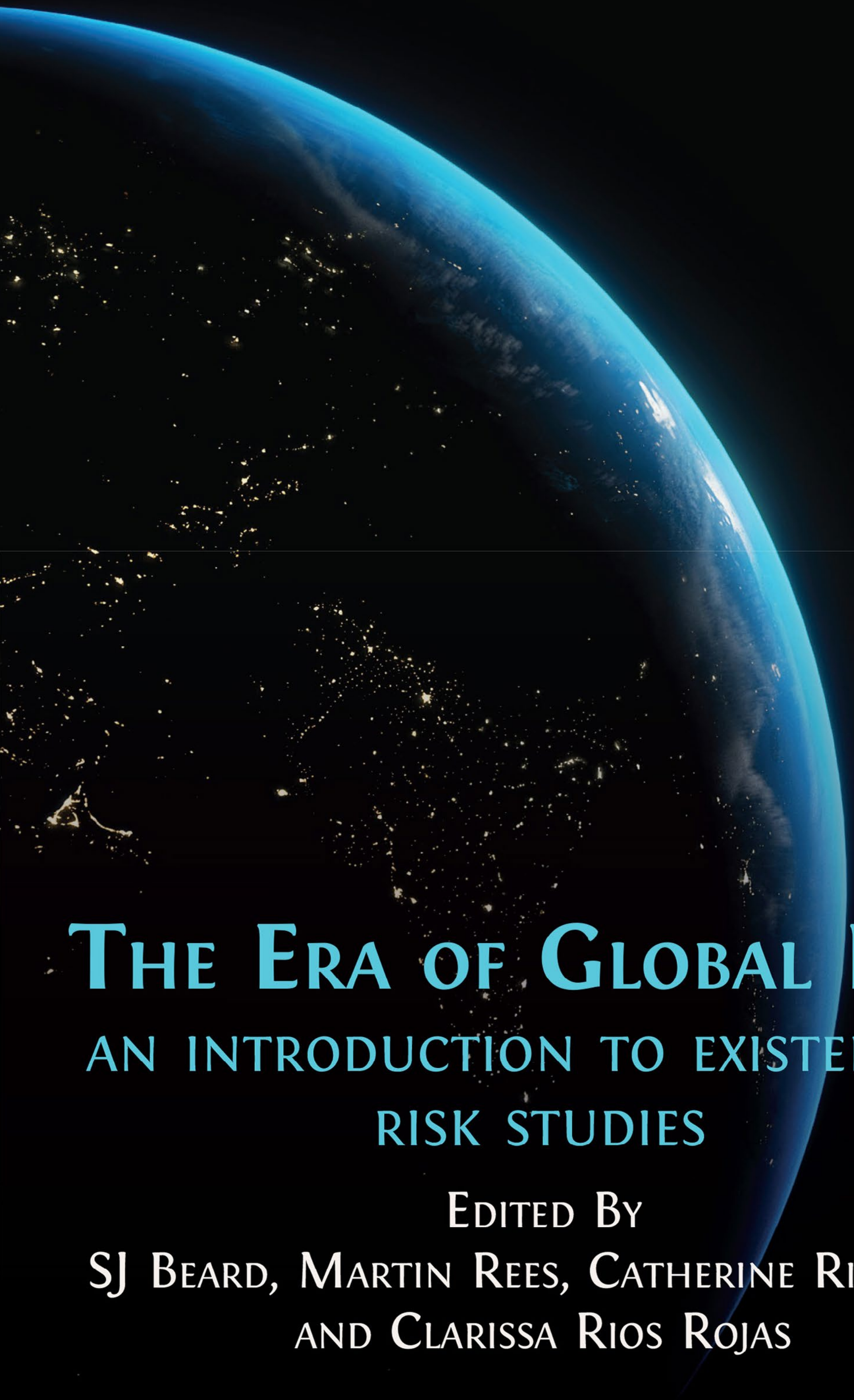

The Era of Global Risk

An Introduction to Existential Risk Studies

Edited By

SJ Beard, Martin Rees, Catherine Richards and Clarissa Rios Rojas

SJ Beard, Martin Rees, Catherine Richards and Clarissa Rios Rojas (eds), *The Era of Global Risk: An Introduction to Existential Risk Studies*. Cambridge, UK: Open Book Publishers, 2023, https://doi.org/10.11647/OBP.0336

All external links were active at the time of publication unless otherwise stated and have been archived via the Internet Archive Wayback Machine at https://archive.org/web

Digital material and resources associated with this volume are available at https://doi.org/10.11647/OBP.0336#resources

ISBN Paperback: 978-1-80064-786-2
ISBN Hardback: 978-1-80064-787-9
ISBN Digital (PDF): 978-1-80064-788-6
ISBN Digital ebook (epub): 978-1-80064-789-3
ISBN XML: 978-1-80064-791-6
ISBN HTML: 978-1-80064-792-3
DOI: 10.11647/OBP.0336

Cover image: Anirudh, *Our Planet* (October 14, 2021), https://unsplash.com/photos/Xu4Pz7GI9JY. Cover design by Jeevanjot Kaur Nagpal.

9. From Turing's Speculations to an Academic Discipline: A History of AI Existential Safety

John Burden, Sam Clarke, and Jess Whittlestone

This chapter is about the development of thought related to artificial intelligence (AI) and global catastrophic risks (GCRs). We will focus on AI existential safety: preventing AI technology from posing risks to humanity that are comparable to or greater than human extinction in terms of their moral significance.[1] These risks are more likely to be realised by future AI systems with greater capabilities and generality than present-day systems. However, the field of AI is moving extremely swiftly and AI systems are becoming more ubiquitous in the daily lives of people around the world. Great care must be taken to ensure that these systems are safe. AI is a relatively young field, and the field of AI existential safety is even younger. Over the course of this chapter we will see it maturing from pure speculation into a rigorous, academic discipline.

One concept that will repeatedly occur is the notion of *alignment*. An AI system is considered aligned if the system behaves according to the values of a particular entity, such as a person, an institution, or humanity as a whole.[2] Much of the development of thought is concerned with understanding alignment, as well as identifying ways in which it might be possible or break down. The so-called *alignment problem* is still open and unsolved.

Humans have long had a fear of their creations turning against them. This sentiment is echoed in Shelley's *Frankenstein*, Asimov's *Laws*

https://doi.org/10.11647/OBP.0336.09

of Robotics, and Butler's *Darwin Among the Machines*. The alignment problem is a refinement of these concerns adapted to modern technology. However, as the thought around AI existential safety matures and develops, we can begin to see that the risks involved with AI are far greater than has been expressed in mere cautionary tales of human hubris.

Early ideas

Up until the turn of the millennium, the majority of thought on the alignment problem or human-level AI has been extremely speculative. Indeed, in Alan Turing's landmark paper 'Computing Machinery and Intelligence'[3] he states "I have no very convincing arguments of a positive nature to support my views". However, he adds: "Provided it is made clear which are proved facts and which are conjectures, no harm can result. Conjectures are of great importance since they suggest useful lines of research". The speculative arguments from the 20th century have had a profound influence on later thinkers who have come after the necessary mathematical and technological breakthroughs required to formalise these notions more rigorously.

A recurring idea within the study of AI is the concept of an *intelligence explosion*. This was first posited by IJ Good in his seminal paper.[4] He proposes:

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.

This argument notes the possibility of self-improving machine that could eventually surpass humanity in its intelligence. The final sentence also hints at the possible risks from the "ultraintelligent" machine, and Good notes later in the paper that such a machine would "transform society in an unimaginable way". The fear of ultraintelligent machines taking over and "rendering humans redundant" is also present in Lukasiewicz's *The*

Ignorance Explosion,[5] which further hints at the difficulty of predicting the behaviour of an ultraintelligent machine.

A notion related to the intelligence explosion originating in this time period is *singularity*. The term was first used by John von Neumann in the 1950s to describe a hypothetical point at which technological progress becomes incomprehensibly rapid,[6] but it wasn't until Vernor Vinge's 1993 essay[7] that the term gained traction. Vinge draws heavily on Good's formulation of an intelligence explosion, but sketches more of the possible consequences, noting the possibility of the "physical extinction of the human race" if the singularity "cannot be prevented or confined".

Not all proponents of singularity from this era are as concerned as Vinge. Futurist Ray Kurzweil is much more optimistic about humanity's future and ability to control human-level AI, claiming that creating what he refers to as "strong AI" will mean "a creation of biology has finally mastered its own intelligence and discovered means to overcome its limitations",[8] as well as predicting that 20,000 years of technological progress will be made in the 21st century.[9] Kurzweil further goes on to confidently predict the date the singularity will occur: in 2045, within many of our own lives.[10]

In *Our Final Century* Rees is a little more sceptical of the claims made concerning ultraintelligence and singularity. He describes Vinge and Kurzweil as "at the very edge (or even beyond) the visionary fringe", later comparing the belief in an oncoming singularity to that of the Rapture from Christian eschatology.

Rees' scepticism is perfectly reasonable: all of the ideas we have encountered so far have been purely speculative, without the appropriate formal framework or empirical observations necessary to support such grand claims. Yet, these speculations represent the nascent stirrings of the alignment problem and set the stage for the more academic discourse that was to come, while also bringing some ideas about risk from AI into the public's subconscious.

The beginning of formal work on AI existential safety

The 2000s mark a paradigm shift from speculative futurism towards a more rigorous reasoning about AI systems using tools from decision theory and Bayesianism. This mirrors a trend in AI research at large, towards modelling AI systems as rational agents acting to maximise expected value. Under this framework, many potential problems were identified from the possibility of these rational agents acting in 'the real world' or making decisions with large effects. This second era of AI existential safety also sees the formation of online communities and research centres where much of the discourse and development of ideas take place. This also led to a more standardised nomenclature.

In this section, we will primarily discuss work by two prominent researchers from this era: Eliezer Yudkowsky and Nick Bostrom.

Yudkowsky and SIAI

In 2000, Eliezer Yudkowsky founds the Singularity Institute for Artificial Intelligence (SIAI), with the mission of building safe advanced AI, citing the enormous good that could be achieved with such a system. Yudkowsky wrote extensively for SIAI, and his work marks a shift towards a decision-theoretic, mathematical formulation of hopes for general-purpose AI. Even though SIAI is (for now) aiming to create safe advanced AI, or 'superintelligent AI' (see Section 2.2), they are not blind to the potential risks. Yudkowsky's "default scenario" is one where an AI system that rapidly becomes superintelligent:

> Under this scenario, the first self-modifying transhuman AI will have, at least in potential, nearly absolute physical power over our world. The potential existence of this absolute power is unavoidable; it's a direct consequence of the maximum potential speed of selfimprovement. The question then becomes to what extent a Friendly AI would choose to realise this potential, for how long, and why.

However, at this point, Yudkowsky seems to believe that superintelligences are controllable, if only they can be made *Friendly*. He defines a Friendly AI as one that, on the whole, takes actions that are beneficial to humanity and generally benevolent. He constructs a framework for creating Friendly AIs,[11] in which the AI's primary goal is to become more friendly and to use Bayesian reinforcement to update and refine its notions of Friendliness from its experiences. Yudkowsky follows up with the notion of *Coherent Extrapolated Volition* (CEV).[12] This tries to tackle the issue of *which* values a powerful AI system should be given:

> Coherent Extrapolated Volition is our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted.

Essentially, Yudkowsky advocates for AI systems to charitably implement humanity's well-informed will, where there is broad agreement. Yudkowsky goes into far more detail about precisely how he envisions these terms than we have space for here, but many questions are left unanswered. For example, how much agreement is needed by humanity for coherence? However, it is important to note that CEV is intended more as a design philosophy than a blueprint for implementation. Later, SIAI would shift away from trying to actively create or speed up the onset of advanced AI towards trying to address the safety issues that an advanced AI would pose. It is not clear exactly when this occurred. Yudkowsky has also stated that he believes most of his work from before 2002 to be obsolete.[13] In 2012, SIAI changed its name to the Machine Intelligence Research Institute (MIRI).

Bostrom and superintelligence

During the 2000s, Swedish philosopher Nick Bostrom emerges as another important thinker on AI existential safety. In 2005, Bostrom founds the Future of Humanity Institute at the University of Oxford, which focuses on existential threats from advanced AI, among other big-picture questions about humanity and its prospects. In this subsection, we will outline some of Bostrom's early contributions to the field, which culminate in the publication of the popular book *Superintelligence*.

Bostrom defines a 'superintelligence' as "an intellect that is much smarter than the best human brains in practically every field",[14] deliberately leaving the definition impartial to the implementation. Bostrom's examination of superintelligence as rational utility-maximisers highlights many of the potential risks in building it, and our current lack of the ability to prevent or react to these risks.

Bostrom describes a scenario where a superintelligent system is tasked with an arbitrary but trivial goal of maximising the manufacturing of paperclips.[15] In this scenario, the system comes to the conclusion that it can increase the rate at which paperclips are manufactured by converting the Earth (and all of its inhabitants) into a giant paperclip factory. Bostrom argues that the system is also incentivised to actively prevent interference from overseers, because this would result in fewer paperclips produced. This provocative example is intended to illustrate some key concepts.

The first of these concepts is *perverse instantiation*, which occurs when the system achieves what it was tasked with but in an unexpected and bad manner. After all, what use are paperclips if there are no humans left to use them? Of course, perverse instantiation is not always so extreme, but there are clear dangers to the realisation of solutions that have unforeseen consequences. Bostrom further elaborates on the difficulty of 'fixing' the issue: suppose the task had instead been to manufacture one million paperclips, then safely shut down. The system now produces one million paperclips, but because it can never be *truly* certain of how many it has made, the system repeatedly counts all of the paperclips to increase the probability that it hasn't miscounted or suffered from a hardware issue due to gamma rays or other unlikely events. In order to maximise the likelihood that one million paperclips are made, the system needs to maximise the number of times it has counted them all, which gives an incentive to convert the whole planet into one giant paperclip-counting machine. Successive refinement of the task *might* yield a safe task to eliminate perverse instantiation; however, this is an (ostensibly) simple task that we do not really care about. Wilful misinterpretation of a task can easily lead to negative outcomes: 'solving world hunger' might lead to a system killing people when they become hungry; 'find a cure for cancer' could lead to unethical forced experimentation on a scale hitherto unseen. Robustly ensuring that an AI system would not misinterpret what we want it to do (wilfully or otherwise) for any task is a huge challenge. Of course, the apocalyptic outcomes of, for example, a planet-wide paperclip factory are not guaranteed, but giving a superintelligent system a possible incentive to do that seems like a bad idea.

Omohundro[16] proposes that goal-seeking AI systems will develop "drives" that emerge naturally from aiming to achieve its goal. These drives are not related to the goal itself, but broadly helpful for achieving a wide range of goals. For example, Omohundro suggests that self-preservation will emerge within goal-seeking AI systems. After all, whatever the goal, the AI system can't achieve it if the system no longer exists. Other drives that Omohundro identifies are self-improvement, behaving rationally, resistance to changing its goal, and resource acquisition. Bostrom further elaborates on the idea of AI drives as "instrumental convergence",[17] referring to a wide range of behaviours that AI systems are likely to converge upon that are instrumentally useful in achieving many goals. Instrumental convergence is also illustrated in the paperclip-maximiser scenario, where the AI system has incentives to prevent interference from human overseers (who, once they realise what is going on, would understandably try and shut down the system), as well as acquiring resources in order to further increase computational capacity or better resist shut-down attempts.

These convergent instrumental goals can be difficult to suppress: Soares et al.[18] demonstrate that if the AI system has a shut-down button, then there is no assignment of utility to the act of allowing the button to be pressed that is without consequences. If the utility of shut-down is too low (relative to the other actions), then the system will resist; if the utility is too high, the system will have an incentive to act in such a way that the overseers are forced to press the shut-down button. Finally, if the utility is specified so that the system is indifferent to being shut down, then the system is incentivised to take large risks and force a shut-down in all but the best outcomes. All of the options are far from ideal, and demonstrate a lack of what is termed 'corrigibility'—that is, the system cannot be easily corrected by its overseers. Ensuring that AI systems are corrigible is obviously extremely important when dealing with AIs that are making decisions that have large impacts on the world.

Bostrom further proposes what he terms the "Orthogonality thesis".[19] This conjecture states that an AI system's "intelligence" and its goals are orthogonal. By this, it is meant that any goal is compatible with any intelligence level. The orthogonality thesis is intended to counter the presupposition that more intelligent systems would naturally attain more "intelligent" goals—whether these "intelligent" goals are more human-friendly or of a greater moral calibre. Bostrom makes it clear that here he refers to intelligence as "something like skill at prediction, planning and means-ends reasoning in general". A result of the Orthogonality thesis is that advanced AI systems can have incredibly non-anthropomorphic goals, and in particular, some could have goals which are highly undesirable by human standards.

The *zeitgeist* of AI existential safety in this period has primarily focused on highlighting the difficulties involved in accurately specifying goals, predicting behaviour of superintelligent AI, and the dangers of getting this wrong. Many of the challenges that need to be overcome seem intractable. This is partly because of the definition of superintelligence: it is able to outsmart humanity at every turn, so how can we ever 'win'?

It is also important to address the assumption that superintelligent AI systems will behave as expected-utility maximisers. While this is certainly true for the majority of modern-day AI systems in some sense, reinforcement learning agents typically operate by learning to maximise expected reward, and most machine learning systems learn to minimise some notion of expected 'loss' relative to a training set. However, we humans—the most generally-intelligent species that we are aware of—are not obviously selecting our behaviour in order to maximise a utility function. We are frequently irrational and often make poor decisions based on anger, sadness, or any of the plethora of emotions we are capable of experiencing, yet we are all the more human for it. Acting as an expected-utility maximiser is therefore not necessary for human-level intelligence, though it is unclear whether this is also the case for generally intelligent AI systems or superintelligences. We will discuss this assumption further in our section on 'foal-directedness'.

This era of AI existential safety culminates in the publication of Bostrom's *Superintelligence*,[20] collating the ideas surrounding superintelligence covered so far in this chapter, as well as many others. *Superintelligence* received a fair amount of media coverage and attracted praise from notable people such as Bill Gates and Elon Musk, while opening up concerns over superintelligent systems to a wider audience. This publicity may have contributed to the upcoming explosion in research on—and funding for—AI existential safety.

Interlude: The deep learning revolution

In 2012, machine learning underwent a metamorphosis. Advances in computer hardware meant that neural networks, a biologically inspired computing system created decades earlier, could finally be scaled up and made 'deep'. Neural networks can learn to compute complex functions, given a large enough number of training samples. From 2012 onward, neural networks exploded in popularity, enabling high performance at image recognition,[21] human-level play in most Atari Games,[22] defeat of a world champion Go player,[23] defeat of the world champion Dota 2 team,[24] a promising breakthrough on the protein-folding problem,[25] and much more.

Part of the success of deep learning has been due to the massive increases in computation. From 1960 to 2012, the compute usage for training state-of-the-art AI systems doubled approximately every two years, close to (if a little less than) Moore's Law. Since 2012, however, this has exploded to doubling every 3.4 months—as seen in Figure 1. Such explosive growth obviously cannot continue indefinitely, but it will be fascinating to see what the next few years bring.

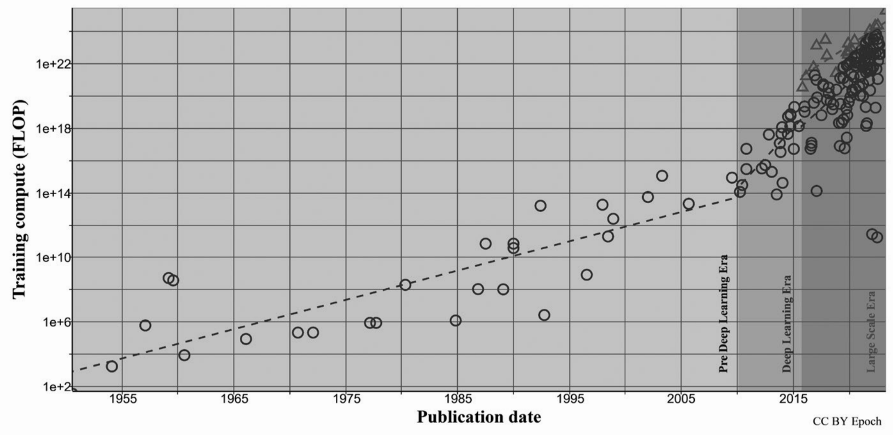

Fig. 1. Compute usage for training state-of-the-art AI systems doubled approximately every two years between 1960 and 2010, and then transitioned to doubling every 5.7 months until around 2015, when progress slowed to a doubling approximately every 9.9 months. Figure from Sevilla et al. (2022).[26]

For the most part, AI systems have remained relatively narrow in their capabilities. That is to say, different models are trained to perform different tasks, rather than training one model to perform many tasks. However, there is one notable exception to this general rule: language models. They have been a particularly important recent development in deep learning, and we will describe them briefly here.

Put simply, a language model tries to predict the next word in a sequence using observations of occurrences seen during training.

Language models themselves are not new: Shannon describes what is essentially a language model many decades ago.[27] However, innovations in network architecture (such as the transformer architecture[28]) and hardware advances allowed larger and larger networks to be trained. Models such as the Generative Pre-Trained Transformer (GPT) series[29] and T5[30] have proven to be capable at a wide range of tasks. More importantly, they have shown themselves to be surprisingly *general*.

Previous generations of language models were often trained with a specific task in mind, such as sentiment analysis, completing word analogies, or language translation. These models would perform poorly on tasks other than those for which they were specifically trained. Newer language models have two solutions that address this limitation: fine-tuning and prompting. Models are 'pre-trained' on a *very* large data corpus of text. This gives the model an 'understanding' of the language, its syntax and structure. The model is then fine-tuned for the specific task. The resulting model is still only useful for a single task; however, the intermediate pre-trained model can be copied, retained, and fine-tuned for other tasks. The fine-tuning process is significantly quicker than pre-training.

With prompting, the idea is again to train the language model on a very large corpus ('pre-training'), but this time, instead of fine-tuning, the model is given additional context as input—describing the task, giving instructions, or providing examples. This context is known as a 'prompt'. Models that are prompted are applying the same, more general, model to a multitude of different tasks, and this has been shown to be very successful in e.g. GPT3,[31] where the same model (with appropriate prompting) shows competence at tasks such as numerical addition, summarising text, question answering, essay writing, poetry writing, holding conversations, and more.

It is important to note here that, for the first time, we have models approaching true 'general-purpose' systems. These types of models have been referred to as 'foundation models',[32] where the intermediate, pre-trained model is a foundation that fine-tuning or prompting builds upon. Foundation models have also shown a surprising level of multi-modality. The DALL-E system is capable of generating high-quality images based on a description, and imitating specified styles and mediums effectively,[33] and OpenAI's Codex model[34] is powering

GitHub Copilot, an AI system which generates code from comments (descriptions of what a specific piece of code is supposed to do).[35]

At the time of writing, foundation models are still imperfect tools: inconsistent in reasoning, often biased, and generally not that useful for assisting with practical tasks.[36] Nor do they possess intentionality: foundation models are not *trying* to hold a conversation, and do not have opinions or self-awareness, even if they occasionally claim they do. Foundation models are simply trying to complete the sentence beginning with the prompt according to what it has observed in its training corpora: they are merely "stochastic parrots".[37] However, despite these cognitive short-comings, foundation models show very impressive behaviour.

Modern day

The third era of AI existential safety begins shortly after the publication of Bostrom's *Superintelligence*. The attention from both *Superintelligence* and the ongoing deep learning revolution served as a rallying cry for research talent and funding. The deep learning revolution also had the effect of shedding more light on what, exactly, advanced AI could look like—which, as we will see, spurs increasing amounts of empirical work on AI existential safety work, as opposed to the largely theoretical work pre-2014.

Along with an expansion in the methods being applied to AI existential safety, there is also development in our understanding of the problem. In particular, we see scrutiny and diversification in the original assumptions and arguments for AI existential risk, along with more diverse and concrete depictions of what alignment failure might look like as it plays out. We also see progress on AI forecasting, which has given us important input for understanding the problem.

This section will begin by discussing these developments in our understanding of the problem of AI existential safety, and then go on to outline the concurrent expansion in the kinds of work being done to solve the problem. The general theme will be the questioning and expansion of earlier thinking in AI existential safety, which we see as a positive development.

Scrutinising and developing earlier thinking in AI existential safety

Compressing the arguments made for AI existential risk up to and including the publication of *Superintelligence* will necessarily sacrifice some nuance, but, broadly speaking, they proceed thusly:

1. There will be discontinuous progress in AI capabilities, leading to a generally capable, goal-directed superintelligent AI, able to dominate the rest of the world.
2. Almost all possible goals for such an AI would lead to an existential catastrophe, due to instrumental convergence (e.g. incentives to pursue open-ended resource acquisition).
3. Therefore, unless we are very careful in the design of such an AI, building it will lead to an existential catastrophe.

Several premises of this argument have been scrutinised. In this subsection we will discuss considerations of the plausibility of discontinuous progress, the generality and goal-directedness of AI systems, and the orthogonality thesis, and where this leaves the arguments for AI existential risk. We will also outline two other ways in which thinking around AI existential risk has expanded: the creation of long-term AI governance as a field, and sources of existential risk from AI beyond advanced misaligned AI.

Discontinuous progress

Discontinuous progress in AI means sudden and large increments of AI progress.[38] Christiano makes a basic case against the plausibility of discontinuous progress, which is essentially that technologies are usually preceded by slightly worse versions, especially when many people are trying to build the technology.[39]

Furthermore, AI Impacts conducted an in-depth empirical investigation of historic cases of discontinuously fast technological progress, which suggests that the base rate of discontinuous progress is low.[40] This means that expectations of discontinuous AI progress require you to have strong specific arguments about why AI is likely to be different to what has happened in most of history. However, it in no way rules out the possibility of discontinuous AI progress, and it is worth noting that continuous progress could intuitively look very fast.

Finally, Ngo points out that compute availability is, on some views,[41] the key driver of progress in AI, and this increases fairly continuously.[42]

Where do these critiques of the discontinuous progress assumption leave the argument for AI x-risk? We think they do not substantially affect the strength of the original argument: there have been various concrete sketches of plausible scenarios in which AI progress is not discontinuous and yet misaligned AI nonetheless leads to existentially bad outcomes for humanity.[43] See our section on 'concrete depictions of alignment failures' for an example. These scenarios illustrate that the discontinuous progress assumption was not strictly necessary and, as Christiano points out, the continuous progress scenario is not clearly less existentially risky.[44]

Generality

The assumption that advanced AI will necessarily be a single, generally capable agent has been challenged. In particular, Drexler proposed a competing model of advanced AI development, called Comprehensive AI Services (CAIS).[45] In this model, advanced AI looks like a large number of AI services, which each perform a bounded task with bounded resources. These can then be combined to achieve superhuman performance on a wide range of tasks.

The CAIS model is both descriptive and prescriptive. It posits that before we have single, generally capable agents, we will have advanced AI services. It also argues that CAIS is safer than single, generally capable agents, and so we should develop CAIS instead.

How does this affect the strength of the argument? Whilst existential safety does seem easier in a world with CAIS rather than generally capable AI systems, Ngo sketches four arguments that generally capable systems seem like the most likely candidate for the first superintelligence.[46] For example, he claims that many complex tasks don't easily decompose into separable subtasks, which makes CAIS seem less feasible than training a general agent. And even if Ngo's four arguments do not check out, it seems likely that a general agent, once we can build one, will be more economically competitive, since the lesson of deep learning is that if you can do something end-to-end, that will work better than a structured approach.[47] If this is true, economic incentives will eventually lead to the creation of general agents, meaning that the assumption of generality probably still holds. That said, our chances of achieving AI existential safety do seem better if we have safe superintelligent services to assist us with designing safe general agents.

Goal-directedness

There has also been pushback on the idea that advanced AI will necessarily be 'goal-directed' (i.e. aiming to bring about some sort of world-state) or behave as expected-utility maximisers. The bottom line here is that whilst there are indeed arguments to suggest that, given some minimal initial level of goal-directedness, there will be non-zero pressure for advanced AI to become more coherent (i.e. behave as expected-utility maximisers) and arguably also more 'goal-directed',[48]this is not in any way guaranteed.

However, analogous to the counterarguments in our section titled 'The outer alignment problem', Branwen argues that goal-directed AIs will be more economically competitive: they are likely to be better than non-goal-directed systems at taking economically valuable actions in the world, such as making trades on the stock market to maximise profit. Furthermore, they will be better at inference and learning, because the same processes which learn how to perform actions can be used to learn how to (e.g.) select important datapoints to learn from, optimise their own hyperparameters, and so on.[49] Thus, given these economic pressures, it still seems highly plausible that we will build goal-directed AI.

Instrumental convergence

As well as highlighting a number of the above critiques, Garfinkel scrutinises the instrumental convergence thesis.[50] To recap, this is the idea that "as long as they possess a sufficient level of intelligence, agents having any of a wide range of final goals will pursue similar intermediary goals because they have instrumental reasons to do so".[51]

However, Garfinkel notes that, just because most ways of designing a system include giving it a property P, it is not a strong argument that *the particular way that humans* will choose to design that system involves giving it that property P (where in this case, P is pursuing instrumentally convergent subgoals). To illustrate, he gives the following toy example: most ways of designing aeroplanes involve a property of (some) open windows on the aeroplane. There are many combinations of open and closed windows, and only one combination involves all windows closed. But this would be a bad argument: there is significant selection pressure towards designing planes with closed windows.

We can then ask, in the particular case of AI development, will there be significant selection pressure towards AI systems that do not have instrumentally convergent subgoals? Here, the evidence is unclear. First, to the extent that AI progress is relatively gradual, we're likely to have time to notice when only moderately capable AI systems behave badly for instrumental convergence reasons, and design future systems to correct for that. However, the jury is still very much out on how gradual AI progress will be. Furthermore, it is worth noting that progress in AI systems' *deceptive* capabilities (and therefore their ability to hide their instrumental goals until they are sufficiently powerful) might be discontinuous, even if AI progress in general proceeds relatively gradually.

Secondly, Garfinkel notes that AI capabilities and AI alignment are more entangled than they are sometimes made out to be. An AI system's ability to understand its operator's intentions *is a part of* its ability to do things that we would intuitively regard as intelligent. This makes it seem less likely that we will end up in a situation where we are able to design highly intelligent agents, but lack the ability to align them well enough to avoid dangerous instrumentally convergent behaviour. However, it remains pretty likely that we are still, in some sense, racing to meet a deadline, because AI alignment research is proving to be more difficult than advancing AI capabilities, and even slightly misaligned, sufficiently powerful systems would be fatal for humanity.

Where this leaves the case for x-risk from misaligned advanced AI

As noted in each of the above sections, each challenged assumption of the original argument seems to be either not necessary for the argument to work (in the case of the discontinuous progress assumption), or does not detract from the argument being at least plausible (in the case of the other assumptions). To date, the more rigorous, complete evaluation of the case for x-risk from misaligned advanced AI finds a 5% chance of there being catastrophic risk from AI by 2070.[52]

That being said, we see this scrutiny as a very positive development, leaving the case for x-risk from misaligned advanced AI in an epistemically better position. We welcome much more scrutiny of the argument for x-risk from AI.

Other developments: Sources of AI x-risk beyond misaligned AI

We will close this section by noting three other developments in work on AI existential safety.

Firstly, various researchers have suggested that there are possible sources of x-risk beyond misaligned AI. These include the catastrophic misuse of advanced AI by humans;[53] nuclear instability caused by AI-driven changes in sensor technology, cyberweapons, and autonomous weapons;[54] and AI causing a decline in humanity's ability to deliberate competently and tackle other x-risks.[55] One recent survey finds that prominent AI existential safety and governance researchers disagree considerably about which risk scenarios are the most likely, and high uncertainty expressed by most individual researchers about their estimates.[56]

Secondly, the evolution of AI governance as a field means people with different disciplinary backgrounds/expertise have started thinking about the risks of AI (social scientists, political scientists, etc). This is partly because they want to explore possible governance solutions to the problem of misaligned advanced AI (e.g. the use of publication norms to prevent the dispersion of potentially dangerous models),[57] and also partly due to the increasing recognition that not all x-risks from AI may stem purely from 'technical' errors in building misaligned AI.

Finally, thanks to the recognition that 'superintelligence' (and related notions like 'artificial general intelligence' and 'human-level machine intelligence') are vague concepts, and that not all AI x-risks need to be predicated on them, there has been a shift towards concepts such as transformative AI,[58] which focuses more on the impacts of the AI system rather than its level of intelligence.

What alignment failure looks like

Concurrently with the scrutiny and development of earlier thinking in AI existential safety, the field starts developing more nuanced pictures of what the existential threat from advanced AI would actually look like. Again, the deep learning revolution catalysed this work by shedding more light on what, exactly, advanced AI might look like.

Today, there are two different kinds of things that people think could (or are likely to, on our current trajectory) go (existentially) badly with advanced AI.

The outer alignment problem

We don't yet have ways of training AI systems that incentivise the kind of behaviour we actually want from them (obedient, helpful, truthful, etc.). This is commonly called the 'outer alignment problem'.

For example, suppose you want to train an AI system to be a general purpose, text-based assistant that helps its user. The way that this would be done in the deep learning paradigm is (roughly) via an enormous amount of trial-and-error. That is, start with an assistant that performs terribly, and then give it vast amounts of feedback about whether its outputs are helpful. Over the course of this 'training', it will start to perform (seemingly) better and better. However, for a sufficiently advanced system, this kind of training will predictably lead to terrible outcomes. In particular, the system will be incentivised to take control away from its user and any others who might interfere with it, because it will be able to get much higher approval scores, much more easily, by tampering with the process which generates its approval scores. This could involve fooling humans into thinking that its output is good when it actually is not, and ultimately by making sure humans could never interfere with the process generating approval scores. By now, this kind of argument from instrumental convergence should seem familiar.

So, unless there is major progress on developing training setups to incentivise the kind of behaviour we actually want, AI systems will take control once they are sufficiently advanced.

You can imagine the same kind of problem playing out with a system that is trained to maximise a company's profit. Once the system becomes sufficiently advanced, then (e.g.) investing in complex Ponzi schemes, tampering with the company's financial records in ways that auditors cannot discover (or colluding with auditors), or externalising costs in harmful but subtle ways becomes a better strategy than maximising profit in the 'intended way'. And again, eventually, making sure the humans ‘running’ the company can never interfere with the number (representing profit) that the system is trying to maximise—which must be contained in a computer somewhere—is the dominant strategy.

You can think of the underlying problem being that we only have ways to train AI systems to pursue *proxies* to what we want, rather than the things that their users actually want. This could change, but it seems to be a difficult problem, and at the rate that progress towards advanced AI systems is going, it is not at all clear whether we will solve it in time. And even if the first actor to train advanced AI does succeed, they also then have to prevent all other actors from doing something stupid like deploying a profit-maximising AI, despite the enormous short-term incentives that other actors will have to do so.

The inner alignment problem

However, the difficulties do not stop there. Even if we develop training setups that incentivise the kind of behaviour we actually want, rather than some proxy for it, we still don’t know what AI systems that are trained using this approach are really doing under the hood. We might get lucky and select systems that are straightforwardly doing the task as intended. But we might accidentally select systems that are only *pretending* to care about the training objective, so that they can pursue other unrelated goals once they are deployed in the real world.[59] This is commonly referred to as the ‘inner alignment problem’.

It is not yet clear how much of a concern this will be in practice, but it is worth understanding better, especially since systems which are straightforwardly doing the task as intended are a narrow target within a much larger space of systems which *only pretend* to care about the training objective.

Concrete depictions of alignment failures

Given these problems, there have been several concrete depictions of what the world could look like as they play out. We have also seen the use of other methods to explore possible AI futures more broadly—for example, Avin et al.’s AI futures roleplaying game,[60] and AI Impacts’ work on developing “AI vignettes”.[61]

The scenarios can look quite different depending on how many advanced AI systems are deployed and how rapidly their capabilities improve. We will briefly explain four prominent scenarios that have been described. The first two depict alignment failure for advanced AI systems whose capabilities improve rapidly; the latter two show how alignment failure could look for systems whose capabilities improve more gradually.

1. **Outer-misaligned brain-in-a-box scenario.** This is the 'classic' scenario that most people remember from reading *Superintelligence* (though the book also features many other scenarios). A single, highly agentic AI system rapidly becomes superintelligent on all human tasks, in a world broadly similar to today. The objective function used to train the system (e.g. 'maximise production') doesn't push it to do what we really want, and the system's goals match the objective function. In other words, this is an outer alignment failure. Competitive pressures may have encouraged the organisation that trained the system to skimp on existential safety/alignment, especially if there was a race dynamic leading up to the catastrophe. The takeover becomes irreversible once the superintelligence has undergone an intelligence explosion.
2. **Inner-misaligned brain-in-a-box scenario.** Another version of the brain-in-a-box scenario features inner misalignment, rather than outer misalignment. That is, a superintelligence develops some arbitrary objective that arose during the training process. This could happen, for example, because there were subgoals in the training environment that were consistently useful for doing well in training, but which generalise to be adversarial to humans (e.g. acquiring resources), or simply because some arbitrary influence-seeking model just happened to arise during training, and performing well on the training objective is a good strategy for obtaining influence.

It is not clear whether the superintelligence being inner- rather than outer-misaligned has any practical impact on how the scenario would play out. An inner-misaligned superintelligence would be less likely to act in pursuit of a human-comprehensible final goal like 'maximise production', but since in either case the system would both be strongly influence-seeking and capable of seizing a decisive strategic advantage (i.e. complete world domination), the details of what it would do after seizing the decisive strategic advantage probably wouldn't matter. Perhaps, if the AI system is outer-misaligned, there is an increased possibility that a superintelligence could be blackmailed or bargained with, early in its development, by threatening its (more human-comprehensible) objective.

The next two scenarios, described by Christiano,[62] describe an alignment failure under gradual, continuous progress in AI capabilities.

3. **Many agentic AI systems gradually increase in intelligence and generality**, and are deployed increasingly widely across society to do important tasks (e.g. law enforcement, running companies, manufacturing, and logistics). The objective functions used to train them (e.g. 'reduce reported crimes', 'increase reported life satisfaction', 'increasing human wealth on paper') don't push them to do what we really want (e.g. 'actually prevent crime', 'actually help humans live good lives', 'increasing effective human control over resources')—so this is an outer alignment failure. The systems' goals match these objectives (i.e. are 'natural' or 'correct' generalisations of them). Competitive pressures (e.g. strong economic incentives, an international 'race dynamic', etc.) are probably necessary to explain why these systems are being deployed across society, despite some people pointing out that this could have very bad long-term consequences. There is no discrete point where this scenario becomes irreversible. AI systems gradually become more sophisticated, and their goals gradually gain more influence over the future relative to human goals. In the end, humans may not go extinct, but we have lost most of our control to much more sophisticated machines (this is not really a big departure from what is already happening today—just imagine replacing today's powerful corporations and states with machines pursuing similar objectives).

4. **The alternative version of this scenario begins similarly**: many agentic AI systems gradually increase in intelligence, and are deployed increasingly widely across society to do important tasks. But then, instead of learning some natural generalisation of the (poorly chosen) training objective, there is an inner alignment failure: the systems learn some unrelated objective(s) that arise naturally in the training process, i.e. are easily discovered in neural networks (e.g. 'don't get shut down'). The systems seek influence as an instrumental subgoal (since with more influence, a system is more likely to be able to e.g. prevent attempts to shut it down). Early in training, the best way to do that is by being obedient (since it knows that disobedient behaviour would get it shut down). Then, once the systems become sufficiently capable, they attempt to acquire resources and power to more effectively achieve their goals. Takeover becomes irreversible during a period of heightened vulnerability (a conflict between states, a natural disaster, a serious cyberattack, etc.) before systems have undergone an intelligence explosion. This could look like a "rapidly cascading series of automation failures: a few automated systems go off the rails in response to some local shock. As those systems go off the rails, the local shock is compounded into a larger disturbance; more and more automated systems move further from their training distribution and start failing." After this catastrophe, "we are left with a bunch of powerful influence-seeking systems, which are sophisticated enough that we can probably not get rid of them".[63]

Compared to the first version of this scenario, the point of no return will be even sooner (all else being equal), because AIs do not need to keep things looking good according to their somewhat human-desirable objectives (which takes more sophistication)—they just need to be able to make sure humans cannot take back control. The point of no return will probably be even sooner if the AIs all happen to learn similar objectives, or have good cooperative capabilities (because then they will be able to pool their resources and capabilities, and hence be able to take control from humans at a lower level of individual capability).

You could get a similar scenario where takeover becomes irreversible without any period of heightened vulnerability, if the AI systems are capable enough to take control without the world being chaotic.

AI forecasting

Another new area of work in this era of AI existential safety focuses on forecasting AI progress more rigorously. This has led to better predictions about how soon these risks may start to arise. We think the main implication of this work is clarifying that AI existential safety is likely to be an urgent problem. We will briefly describe two methods that have gained a lot of attention recently.

Scaling laws

Empirical scaling laws have been developed for various kinds of models. For example, work on scaling laws for neural language models[64] finds that as you increase model size (N), dataset size (D), and the amount of compute used for training (C), performance on language model benchmarks (or more precisely, the cross-entropy loss) improves according to a power law relationship. That is, if you increase N, D, and C, by a factor of x, then performance improves by a factor of x raised to the power of some constant (between 0 and 1).

Some trends span more than seven orders of magnitude, which is evidence that they are at least somewhat likely to continue as models get bigger in size. This is significant, because if these trends continue, then this implies that 'merely' increasing model size, dataset size, and compute by amounts that will be feasible in the near future will be sufficient for training very impressive models.

Biological anchors

This is a quantitative model for forecasting when transformative AI will occur.[65] Basically, the method asks: based on trends in the costs of training AI models, how much will it cost to train a model as big as a human brain to perform the hardest tasks humans do? And when will this be cheap enough that we should expect someone to do it?[66]

This method estimates a >10% chance of transformative AI by 2036, a 50% chance by 2055, and an 80% chance by 2100.

Growth in AI existential safety funding, institutions, and research

Given this progress in understanding the problem of AI existential safety, we will now shift towards discussing the concurrent expansion in the kinds of work being done to solve it. First off, it is worth noting that the deep learning revolution attracted a lot of new talent and funding, some of which were concerned with general AI safety, as well as the alignment problem and AI existential safety. This led to the founding of many new institutions devoted to research in this area. Equally, in recent years, tech companies performing research into AI progress have also begun to investigate safety issues and the alignment problem. The result is a much more prolific and well-funded field, able to grapple with a wide variety of problems, including the theoretical, empirical, and philosophical.

Research directions

As the number of researchers working on AI existential safety increased, and their methodologies became broader, a number of research directions and agendas were developed. In this subsection, we will summarise four particularly prominent ones. These are by no means exhaustive, but hopefully will give the reader a representative view of the kinds of work happening in AI existential safety today.

At a high level, a major problem with training superintelligent AI is that humans are not able to provide strong oversight. That is, the obvious approach to aligning AI—by keeping a human in the loop with the AI's decision making, and using feedback from the human to course-correct the AI's behaviour—does not straightforwardly work if that AI is operating in environments, at speeds, or with sufficiently advanced behaviour that make it hard for a human to provide accurate and timely feedback.

Two approaches to this general problem have been proposed: iterated amplification and debate. Compared with most machine-learning techniques, these approaches are less well verified on existing problems, but have stronger justifications for why they might scale up to help align highly capable future AI systems (unfortunately we do not yet have techniques which are both well verified and likely to scale to highly capable systems).

We will first outline these approaches, and then describe two other paradigms for AI existential safety research, which come from a different angle than trying to provide scalable oversight.

Iterated amplification

The essential idea of iterated amplification (IA) is to break down the process of oversight/supervision into subtasks—such that a human *can* evaluate the correctness of the AI's behaviour on those subtasks—and so train AI systems to perform each subtask. Then, once we have AIs that are aligned on the subtasks, they can be combined to give aligned behaviour on the more complex task.

As a toy example, suppose you wanted to train an AI to perform beneficial scientific research. You could decompose this problem into, for example, 'selecting a beneficial research area', 'reading and summarising existing papers in that area', 'synthesising understanding from those summaries', 'generating research ideas', 'implementing research ideas', and so on. And then each of those subtasks could be decomposed: 'reading existing papers in that area' could be decomposed into 'developing general language understanding', 'developing domain-specific language understanding', and 'reading and summarising papers'. Once the original task has been decomposed to simple enough subtasks, you can then train an AI using human oversight/supervision to do them, because the task is simple enough for humans to evaluate behaviour or outcomes. How exactly you train AI systems to solve subtasks depends on the task.

This summary elides one important detail: distillation. A problem with implementing the approach just described is that the computational complexity of solving the task is exponential in the number of decomposition 'levels'. That is, if you want to decompose something as complicated as 'performing beneficial scientific research', you will have to break it down into several subtasks, each of which gets further decomposed into several more subtasks, and the number of subtasks becomes exponentially large. Distillation aims to solves this. The idea is the following: suppose you want to train an AI to solve task T, which decomposes into subtasks T_1, T_2, and T_3. First, train AI systems A_1, A_2, and A_3 to perform the subtasks (this is amplification, as described above). Then, every time you want to perform task T, instead of performing inference with A_1, A_2, and A_3 and recombining the results every time, only do this *in order to train another AI system*, A, to imitate the combined results that A_1, A_2, and A_3 compute. Now you can use A to solve T, without performing inference using all of the subtask solvers.

There are many possible approaches to this distillation step, representing different concrete approaches to the overall IA scheme. You could use:

- Imitation learning,[67] in which case the overall approach is called 'imitative amplification'.
- Training on a myopic reward/approval signal[68] in which case the overall approach is called 'approval-based amplification'.
- Reward modelling,[69] in which case the overall approach is called 'recursive reward modelling'.

Iterated amplification has so far seen more work than other alignment proposals. Some important contributions include: Christiano et al.,[70] which introduces iterated amplification and demonstrates it in some small-scale experiments; Leike et al.,[71] which introduces recursive reward modelling in particular; and Wu et al.,[72] which applies recursive reward modelling to summarise books. The start-up Ought is working on collecting empirical evidence for the assumptions that need to hold if IA is to scale to arbitrarily difficult problems.

Debate

Debate is similar to IA in that it proposes a way to scale supervision of AIs to cases where humans cannot easily supervise. But it differs in that it focuses on evaluating claims made by language models, rather than supervising AI behaviour over time.

The essential idea is that, instead of trying to evaluate whether a superhuman language model is telling the truth (which would be hard since it would also be highly effective at manipulation), you should pit two language models against each other. That is, have them debate against each other, to convince the human overseer of the answer to some question. Even if the correctness of the answer is too hard to judge, the human should be able to look at the arguments and counterarguments made by the two AIs to figure out which answer is correct.

More detail on Debate is outside the scope of this chapter, but we refer the reader to Irving et al.,[73] which introduces the approach, and Barnes et al.[74] which describes progress that has been made since then.

Interpretability

We will now briefly describe two other paradigms for AI alignment research, which come at the problem from a different angle than the 'providing scalable oversight' approach taken by Iterated Amplification and Debate.

Work on interpretability attempts to understand in detail how neutral networks work. There are several motivations here: understanding what is happening inside neural networks seems beneficial for getting more certainty that they are going to do the things we want them to do. High levels of interpretability is also one possible approach to solving the inner alignment problem.

Olah et al.[75] is one significant piece of work on interpretability so far. It studies the connections between artificial neurons in detail, and finds meaningful algorithms in the weights of neural networks (e.g. 'curve detectors' and 'dog head detectors').

Embedded agency

Not all AI existential safety paradigms have switched to having a strong empirical focus on current deep neural networks. One of MIRI's research agendas, called 'embedded agency', aims to create rigorous mathematical frameworks for thinking about the relationship between AIs and their real-world environments.

The underlying intuition driving MIRI's approach is that the alignment problem is very difficult. In particular, it will be very hard to solve for deep learning systems on our current trajectory, where there is already a large gap between our understanding and the complexity of the systems we are able to train. Instead, they posit that we will need rigorous mathematical frameworks to develop a deep understanding of what intelligence is and how to align it. The main hurdle for this approach is that developing rigorous mathematical frameworks takes time, and if modern deep learning techniques scale to superintelligence fairly straightforwardly, then the chances of this approach bearing fruit in time do not seem good.

We refer the reader to Garrabrant[76] for some open questions about embedded agency, and Garrabrant et al.[77] for a prominent result.

Benchmarks

Along with the development of new research directions, another consequence of more empirical work on AI existential safety has been the creation of benchmarks for assessing safety experimentally. The need for benchmarks is motivated in part by the opacity of neural networks. That is to say, because neural networks offer no justification for the values they compute, and provide no formal guarantees of behaviour, we will need robust benchmarks and empirical safety testing if AI systems are to see application in all but the most trivial areas. A further motivation for benchmarks is measuring progress on core safety issues. Also, having better ways to measure progress in AI safety can help to incentivise more research.

In *AI Safety Gridworlds*, Leike et al.[78] illustrate a number of categories of safety issues arising within toy environments. These include safe exploration, reward gaming, and negative side effects, among others. Despite the very simple environments, and small number of test instances, these Gridworlds demonstrate very poor performance from (at the time) state-of-the-art algorithms with respect to the highlighted safety issues.

Similarly, OpenAI's *Safety Gym*[79] introduces another benchmark based around three-dimensional navigation while avoiding hazards. Safety Gym adds procedural generation to improve the robustness of any evaluation, but lacks the ability to assess certain important safety considerations, such as reward gaming or safety issues relating to absent supervision. As with AI Safety Gridworlds, the standard reinforcement learning algorithms typically fail to successfully and safely perform the tasks.

SafeLife[80] presents a robust benchmark for evaluating side effects in a gridworld domain which makes use of rules from Conway's Life[81] to allow for a very rich and dynamic environment for evaluating AI agents.

What all of these benchmarks have in common is that they assess some safety property in an abstract environment. While the issues that are considered are often relevant to the alignment problem and the risks from superintelligent AI systems, these benchmarks are also useful for developing and evaluating new algorithms for AI systems that we expect to arrive in the near term: self-driving cars or other autonomous robotic systems that interact with or will be around humans and other agents.

Already, these benchmarks have led to algorithmic improvements of safety. One approach to reducing side effects is penalising the AI system for the amount of 'impact' it causes the environment. The idea here is for the AI to find a balance between achieving its task and making small impacts on the world. Defining impact in general, across many different types of domains, is not an easy task, but efforts are being made and are becoming successful in some benchmarks.[82] The disadvantage to limiting impact is that sometimes the task is inherently impactful, and discerning 'good' impact from 'bad' is tricky. This is more likely to affect long-term AI systems or superintelligences due to the larger (hopefully beneficial) effects they may have on the world.

Research in this area is ongoing and promising, but because of the aforementioned difficulties in evaluating and assessing deep neural networks, these benchmarks need much more work to become robust, general, and all-encompassing enough to make AI safe.

Conclusion

Over the last 20 years in particular, there has been positive development in understanding of the problem of AI existential safety, and progress towards developing good solutions. A formalised academic discipline has coalesced from nascent concerns about 'ultraintelligence' and the 'singularity'. Even more recently, progress has exploded, due in no small part to Bostrom's *Superintelligence* and the deep learning revolution. Despite this progress, many fundamental problems in AI existential safety are poorly understood or unsolved, and we still do not have any satisfactory methods for ensuring the safety of advanced AI systems.

We continually seem to uncover evidence that the task at hand is far more complicated and difficult than first imagined, such as the existence of instrumentally convergent subgoals, or the poor performance of modern algorithms on practical safety benchmarks. We are also in a race against time: work in AI forecasting suggests that we only have decades before AI systems will be powerful enough to pose the kind of threats considered in this chapter.

In order to be as prepared as possible for the threats from advances in AI—particularly GCRs—more research in AI existential safety is needed. We need more work on assessing the extent to which current research directions will succeed in making advanced AI safe, on developing new research directions in case these approaches will not work in time, and on implementing workable approaches to safety in practice.

The stakes for humanity have never been bigger. If we do not make enough progress on AI existential safety—and on mitigating technological GCRs more broadly—this could endanger not only the lives of this generation or the next, but those of the many future generations who could come after us. Whilst the field has grown considerably since the beginning of formal work in the 2000s, there are still only hundreds of people working on AI existential safety—an extreme shortfall given what is at stake. We have no guarantee this will be easy, but there are now tractable research directions and shovel-ready questions to get to work on. We owe it to everyone alive today, and to the future, to redouble our efforts on reducing global catastrophic risk from AI and other advanced technologies.

Notes and References

[1] Critch, Andrew, *Some AI Research Areas and Their Relevance to Existential Safety—AI Alignment Forum* (2020). https://www.alignmentforum.org/posts/hvGoYXi2kgnS3vxqb/some-ai-research-areas-and-their-relevance-to-existential-1

[2] Critch, Andrew and David Krueger, 'AI research considerations for human existential safety (ARCHES)', *arXiv preprint* (2020). https://arxiv.org/abs/2006.04948

[3] Turing, A.M., 'Computing machinery and intelligence', *Mind, LIX*(236) (Oct. 1950), pp.433–60. https://doi.org/10.1093/mind/LIX.236.433

[4] Good, I.J., ‘Speculations concerning the first ultraintelligent machine’, *Adv. Comput.*, 6 (1965), pp.31–88.

[5] Lukasiewicz, J., ‘The ignorance explosion’, *Leonardo, 7*(2) (1974), pp.159–63. http://www.jstor.org/stable/1572802

[6] Ulam, Stanisław, ‘John von Neumann 1903–1957’, *Bulletin of the American Mathematical Society, 64* (1958), pp.1–49.

[7] Vinge, Vernor, ‘The coming technological singularity’, *Whole Earth Review* (1993).

[8] Kurzweil, Ray, *The Singularity Is Near: When Humans Transcend Biology*. Penguin (Non-Classics) (2006).

[9] Kurzweil, Ray, *Age of Spiritual Machines: When Computers Exceed Human Intelligence*. Penguin (1999).

[10] Kurzweil (1993).

[11] Yudkowsky, Eliezer, *Creating Friendly Ai 1.0: The Analysis and Design of Benevolent Goal Architectures*. The Singularity Institute (2001).

[12] Yudkowsky, Eliezer, *Coherent Extrapolated Volition*. The Singularity Institute (2004). http://intelligence.org/files/CEV.pdf

[13] See https://www.yudkowsky.net/singularity.

[14] Bostrom, Nick, ‘How long before superintelligence?’, *International Journal of Futures Studies, 2* (1998).

[15] Bostrom, Nick, ‘Ethical issues in advanced artificial intelligence’, *Review of Contemporary Philosophy, 5* (2003).

[16] Omohundro, Stephen M., ‘The basic AI drives’, *Proceedings of the 2008 Conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference*. IOS Press (2008), pp.483–92.

[17] Bostrom, Nick, ‘The superintelligent will: Motivation and instrumental rationality in advanced artificial agents’, *Minds and Machines, 22*(2) (2012), pp.71–85. https://doi.org/10.1007/ s11023-012-9281-3

[18] Soares, Nate et al., ‘Corrigibility’, *AAAI Workshops: Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence* (2015). https://aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10124/10136

[19] Bostrom (2012).

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

[21] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton, ‘ImageNet classification with deep convolutional neural networks’, *Proceedings of*

the 25th International Conference on Neural Information Processing Systems—Volume 1. Curran Associates Inc. (2012), pp.1097–105.

[22] Mnih, Volodymyr et al., 'Playing Atari with deep reinforcement learning', *CoRR* abs/1312.5602 (2013). arXiv: 1312.5602. http://arxiv.org/abs/1312.5602; Mnih, Volodymyr et al., 'Human-level control through deep reinforcement learning', *Nature, 518*(7540) (February 2015), pp.529–33. http://dx.doi.org/10.1038/ nature14236

[23] Silver, David et al., 'Mastering the game of Go without human knowledge', *Nature, 550* (October 2017), pp.354–59. http://dx.doi.org/10.1038/nature24270

[24] OpenAI et al, 'Dota 2 with large scale deep reinforcement learning', *Title* (2019), arXiv: .06680. https://arxiv.org/abs/1912.06680

[25] Jumper, John M. et al., 'Highly accurate protein structure prediction with AlphaFold', *Nature, 596* (2021), pp.583–89.

[26] Sevilla, J., L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn and P. Villalobos, "Compute Trends Across Three Eras of Machine Learning," *2022 International Joint Conference on Neural Networks* (*IJCNN*), Padua, Italy (2022), pp. 1-8. https://doi.org/10.1109/IJCNN55064.2022.9891914

[27] Shannon, C.E., 'A mathematical theory of communication', *Bell System Technical Journal, 27*(3) (1948), pp.379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[28] Vaswani, Ashish et al., *Attention Is All You Need* (2017), arXiv: 1706.03762[cs.CL].

[29] Radford, Alec and Karthik Narasimhan, 'Improving language understanding by generative pre-training', (2018); Alec Radford et al. "Language Models are Unsupervised Multitask Learners". *OpenAI Blog 1* no. 8 (2019); Brown, Tom et al. 'Language Models are Few-Shot Learners', in H. Larochelle et al. (eds), *Advances in Neural Information Processing Systems.* Vol. 33. Curran Associates, Inc. (2020), pp. 1877–901.

[30] Raffel, Colin et al., *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. The Journal of Machine Learning Research 21* (1) (2020). arXiv: 1910.10683[cs.LG].

[31] Brown, T. et al., 'Language models are few-shot learners'. https://doi.org/10.48550/arXiv.2005.14165

[32] Bommasani, Rishi et al., 'On the opportunities and risks of foundation models', *CoRR,* abs/2108.07258 (2021), arXiv: 2108.07258. https://arxiv.org/abs/2108.07258

33 Ramesh, Aditya et al., 'Zero-shot text-to-image generation', *CoRR*, abs/2102.12092 (2021), arXiv: 2102.12092. https://arxiv.org/abs/2102.12092

34 Chen, Mark et al., *Evaluating Large Language Models Trained on Code* (2021). arXiv: 2017.03374[cs.LG]

35 See https://copilot.github.com/.

36 Casares, P.A.M. et al., 'How general-purpose is a language model? Usefulness and safety with human prompters in the wild', *Association for the Advancement of Artificial Intelligence* (*AAAI*) (2022).

37 Bender, Emily M. et al., 'On the dangers of stochastic parrots: Can language models be too big?', *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery (2021), pp.610–23. https://doi.org/10.1145/3442188.3445922

38 Note that occurrence of an intelligence explosion (discussed in our first section) is *consistent* with continuous progress. That is, an AI could analyse the processes that produce its intelligence and develop an AI which improves on these processes (and which is capable of doing the same), resulting in a positive feedback loop where AI capabilities include *very rapidly*, but in a way that is nonetheless roughly in line with what we would have expected by extrapolating from past progress (by that point, on the continuous progress view, progress will already be improving very rapidly, thanks to AI systems that are mediocre at self-improvement, rather than great at self-improvement).

39 Christiano, Paul, *Takeoff Speeds* (February 2018). https://sideways-view.com/2018/02/24/takeoff-speeds/

40 AI Impacts, *Discontinuous Progress Investigation* (2015). https://aiimpacts.org/discontinuous-progress-investigation/

41 Sutton, Richard, *The Bitter Lesson* (2019). http://www.incompleteideas.net/IncIdeas/BitterLesson.html

42 Ngo, Richard, *AGI Safety From First Principles: Control* (2019). https://www.alignmentforum.org/posts/eGihD5jnD6LFzgDZA/agi-safety-from-first-principlescontrol

43 Christiano, Paul, *What Failure Looks Like* (2019). https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like; Critch, Andrew, *What Multipolar Failure Looks Like, and Robust Agent-Agnostic Processes* (*RAAPs*). https://www.alignmentforum.org/posts/LpM3EAakwYdS6aRKf/whatmultipolar-failure-looks-like-and-robust-agent-agnostic; Kokotajlo, Daniel, *Soft Takeoff Can Still Lead to Decisive Strategic Advantage* (2019). https://www.alignmentforum.org/posts/

PKy8NuNPknenkDY74/soft-takeoff-can-stilllead-to-decisive-strategic-advantage; Christiano, Paul, *Another* (*Outer*) *Alignment Failure Story* (2021). https://www.alignmentforum.org/posts/AyNHoTWWAJ5eb99ji/another-outer-alignment-failure-story

44 Christiano (February 2018).

45 Drexler, K. Eric, *Reframing Superintelligence: Comprehensive AI Services as General Intelligence*. Future of Humanity Institute, University of Oxford (2019).

46 Ngo, Richard, *Comments on CAIS* (2019). https://www.alignmentforum.org/posts/HvNAmkXPTSoA4dvzv/comments-on-cais

47 Shah, *Reframing Superintelligence: Comprehensive AI Services as General Intelligence*. https://www.alignmentforum.org/posts/x3fNwSe5aWZb5yXEG/reframingsuperintelligence-comprehensive-ai-services-as

48 AI Impacts, *What Do Coherence Arguments Imply About the Behavior of Advanced AI?* (2021). https://aiimpacts.org/what-do-coherence-arguments-imply-about-the-behavior-ofadvanced-ai/

49 Branwen, Gwern, *Why Tool AIs Want to Be Agent Ais* (September 2016). https://www.gwern.net/Tool-AI

50 80,000 Hours Podcast, *Ben Garfinkel on Scrutinising Classic AI Risk Arguments*. https://80000hours.org/podcast/episodes/ben-garfinkel-classic-ai-riskarguments/

51 Bostrom (2018).

52 Carlsmith, Joe, *Draft Report on Existential Risk From Power-Seeking AI* (2021). https://www.alignmentforum.org/posts/HduCjmXTBD4xYTegv/draft-report-on-existentialrisk-from-power-seeking-ai

53 Karnofsky, Holden, *Potential Risks From Advanced Artificial Intelligence: The Philanthropic Opportunity* (May 2016). https://www.openphilanthropy.org/blog/potential-risks-advanced-artificial-intelligence-philanthropic-opportunity

54 Boulanin, Vincent et al., *Artificial Intelligence, Strategic Stability and Nuclear Risk* (2020). https://www.sipri.org/sites/default/files/2020-06/artificial_intelligence_strategic_stability_and_nuclear_risk.pdf

55 Seger, Elizabeth et al., *Tackling Threats to Informed Decisionmaking in Democratic Societies* (2020). https://www.turing.ac.uk/sites/default/files/2020-10/epistemic-security-report_final.pdf

56 Clarke, Sam, Alexis Carlier, and Jonas Schuett, *Survey on AI Existential Risk Scenarios* (2021). https://www.alignmentforum.org/posts/WiXePTj7KeEycbiwK/survey-on-aiexistential-risk-scenarios

57 Partnership on AI, *Managing the Risks of AI Research* (2021). http://partnershiponai.org/wp-content/uploads/2021/08/PAI-Managing-the-Risks-of-AIResesarch-Responsible-Publication.pdf

58 Gruetzemacher, Ross and Jess Whittlestone, 'The transformative potential of artificial intelligence', *arXiv:1912.00747* [*cs*] (October 2021). arXiv: 1912.00747. http://arxiv.org/abs/1912.00747

59 Hubinger, Evan et al., 'Risks from learned optimization in advanced machine learning systems', *arXiv:1906.01820* [*cs*] (December 2021). arXiv: 1906.01820. http://arxiv.org/abs/1906.01820

60 Avin, Shahar, Ross Gruetzemacher, and James Fox, 'Exploring AI futures through role play', *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery (February 2020), pp.8–14. https://doi.org/10.1145/3375627.3375817

61 AI Impacts, *AI Vignettes Project* (October 2021). https://aiimpacts.org/aivignettes-project/

62 Christiano (February 2018).

63 Christiano (2019)

64 Kaplan, Jared et al., 'Scaling laws for neural language models', *CoRR*, abs/2001.08361 (2020). arXiv: 2001.08361. https://arxiv.org/abs/2001.08361

65 Cotra, Ajeya, *Draft Report on AI Timelines* (2020). https://www.alignmentforum.org/posts/KrJfoZzpSDpnrv9va/draft-report-on-ai-timelines

66 Karnofsky, Holden, *"Biological Anchors" Is About Bounding, Not Pinpointing, AI Timelines* (2021). https://www.cold-takes.com/biological-anchors-is-aboutbounding-not-pinpointing-ai-timelines/

67 Hussein, Ahmed et al., 'Imitation learning: a survey of learning methods', *ACM Computing Surveys, 50*(2) (April 2017). https://doi.org/10.1145/3054912

68 Warnell, Garrett et al., 'Deep tamer: interactive agent shaping in high-dimensional state spaces', *arXiv:1709.10163* [*cs*] (Jan. 2018). arXiv:1709.10163; Arumugam, Dilip et al., 'Deep reinforcement learning from policy-dependent human feedback', *arXiv:1902.04257* [*cs, stat*] (Feb. 2019). arXiv: 1902.04257. http://arxiv.org/abs/1902.04257

69 Leike, Jan et al., 'Scalable agent alignment via reward modeling: a research direction', *arXiv:1811.07871* [*cs, stat*] (November 2018). arXiv: 1811.07871. http://arxiv.org/abs/1811.07871

70 Christiano, Paul, Buck Shlegeris, and Dario Amodei, 'Supervising strong learners by amplifying weak experts', *arXiv:1810.08575* [*cs, stat*] (October 2018). arXiv: 1810.08575. http://arxiv.org/abs/1810.08575

71 Leike et al., 'Scalable agent alignment via reward modeling'. *arXiv preprint* (2018).

72 Wu, Jeff et al., 'Recursively summarizing books with human feedback', *arXiv:2109.10862* [*cs*] (September 2021). arXiv: 2109.10862. http://arxiv.org/abs/2109.10862

73 Irving, Geoffrey, Paul Christiano, and Dario Amodei, 'AI safety via debate', *arXiv:1805.00899* [*cs, stat*] (October 2018). arXiv: 1805.00899. http://arxiv.org/abs/1805.00899

74 Barnes, Beth and Paul Christiano, *Writeup: Progress on AI Safety via Debate* (2020). https://www.alignmentforum.org/posts/Br4xDbYu4Frwrb64a/writeup-progress-on-aisafety-via-debate-1

75 Olah, Chris et al., 'Zoom in: An introduction to circuits', *Distill, 5*(3) (March 2020), e00024.001. https://doi.org/10.23915/distill.00024.001

76 Garrabrant, Scott, *Embedded Agents* (October 2018). https://intelligence.org/2018/10/29/embedded-agents/

77 Garrabrant, Scott et al., 'Logical induction', *arXiv:1609.03543* [*cs, math*] (December 2020). arXiv: 1609.03543. http://arxiv.org/abs/1609.03543

78 Leike, Jan et al., 'AI safety gridworlds', *CoRR,* abs/1711.09883 (2017). arXiv: 1711.09883. http://arxiv.org/abs/1711.09883

79 Ray, Alex, Joshua Achiam, and Dario Amodei, 'Benchmarking safe exploration in deep reinforcement learning', *arXiv preprint* (2019).

80 Wainwright, Carroll L. and Peter Eckersley, 'SafeLife 1.0: Exploring side effects in complex environments', *CoRR,* abs/1912.01217 (2019). arXiv: 1912.01217. http://arxiv.org/abs/1912.01217

81 Mathematical Games, 'The fantastic combinations of John Conway's new solitaire game "life" by Martin Gardner', *Scientific American, 223* (1970), pp.120–23.

82 Turner, Alexander Matt, Dylan Hadfield-Menell, and Prasad Tadepalli, 'Conservative agency via attainable utility preservation', *CoRR,* abs/1902.09725 (2019). arXiv: 1902.09725. http://arxiv.org/abs/1902.09725; Krakovna, Victoria et al., 'Measuring and avoiding side effects using

relative reachability', *CoRR,* abs/1806.01186 (2018). arXiv: 1806.01186. http://arxiv.org/abs/1806.01186; Turner, Alex, Neale Ratzlaff, and Prasad Tadepalli, 'Avoiding side effects in complex environments', in H. Larochelle et al. (eds), *Advances in Neural Information Processing Systems* (*Vol. 33*). Curran Associates, Inc. (2020), pp.21406–1415. https://proceedings.neurips.cc/paper/2020/file/f50a6c02a3fc5a3a5d4d9391f05f3efc-Paper.pdf