

Against the singularity hypothesis

David Thorstad (Global Priorities Institute, University of Oxford)

Global Priorities Institute | November 2022

GPI Working Paper No. 19-2022



Against the singularity hypothesis

Abstract

The *singularity hypothesis* is a radical hypothesis about the future of artificial intelligence on which self-improving artificial agents will quickly become orders of magnitude more intelligent than the average human. Despite the ambitiousness of its claims, the singularity hypothesis has been defended at length by leading philosophers and artificial intelligence researchers. In this paper, I argue that the singularity hypothesis rests on implausible growth assumptions. I show how leading philosophical defenses of the singularity hypothesis (Chalmers 2010, Bostrom 2014) fail to overcome the case for skepticism. I conclude by drawing out philosophical implications of this discussion for our understanding of consciousness, personal identity, digital minds, existential risk, and ethical longtermism.

1 Introduction

The Association for the Advancement of Artificial Intelligence recently convened a panel to discuss the long-term future of artificial intelligence. That panel lamented “a tendency . . . to dwell on *radical* long-term outcomes of AI research, while overlooking the broad spectrum of opportunities and challenges” raised by machine intelligence (Horvitz and Selman 2012, p. 302). Rather than focusing on extreme scenarios involving murderous androids or superhuman planners, we might concentrate on better-evidenced problems such as transparent decisionmaking (Bellamy et al. 2019; Lipton 2018) and labor market distortions (Korinek and Stiglitz 2019).

One hypothesis in particular drew the panel’s ire. The *singularity hypothesis* begins with the supposition that artificial agents will gain the ability to improve their own intelligence. From there, it is claimed that the intelligence of artificial agents will grow at a rapidly accelerating rate, producing an *intelligence explosion* in which artificial agents quickly become orders of magnitude more intelligent than their human creators. The result will be a singularity, understood as a fundamental discontinuity in human history beyond which our fate depends largely on how we interact with artificial agents.

Despite the ambitiousness of its claims, the singularity hypothesis has found no shortage of advocates. The singularity hypothesis has been defended at length by philosophers such as David Chalmers (2010; 2012), Nick Bostrom (2014) and I.J. Good (1966). Similar defenses have been penned by foundational figures in the history of artificial intelligence, including Ray Solomonoff (1985) and Stuart Russell (2019). The singularity hypothesis is endorsed by a nontrivial minority of contemporary artificial intelligence researchers.¹ It

¹For example, Grace et al. (2016) polled authors of accepted papers at two prestigious machine learning conferences, finding a median 10% probability of superintelligence being developed within two years of human-level intelligence, with over half of respondents assigning greater than 40% probability to a weakened form of the intelligence explosion argument. Preliminary results from Stein-Perlman et al. (2022) replicate this analysis. See also Baum et al. (2011), Müller and Bostrom (2016) and Zhang et al. (2022).

has been the subject of a special issue of the *Journal of Consciousness Studies* (Awret 2012) and several volumes of essays (Eden et al. 2012; Callaghan et al. 2017). The singularity hypothesis has become intertwined with philosophical debates about consciousness (Dennett 2012), digital minds (Dreyfus 2012), and philanthropic giving (Greaves et al. forthcoming). And the hypothesis attracts continued interdisciplinary discussion (Armstrong et al. 2016; Yampolskiy 2016).

These efforts have had a major impact on societal conceptions of artificial intelligence. The singularity hypothesis entered the public consciousness through popular works by Vernor Vinge (1993) and Ray Kurzweil (2005). It has been a primary subject of research at institutes such as the Singularity Institute for Artificial Intelligence and the Singularity University (Yudkowsky 2013).² And we will see in Section 6 that the singularity hypothesis had a significant effect on philanthropy through the effective altruism movement, contributing to a large shift of philanthropic giving away from causes such as global health and development aid towards *longtermist* causes such as the prevention of existential catastrophe from self-improving artificial intelligence (MacAskill 2022; Greaves et al. forthcoming; Ord 2020).

Despite ongoing academic and public support for the singularity hypothesis, I argue that the case for the singularity hypothesis is surprisingly thin. Here is the plan. Section 2 works towards a rigorous statement of the singularity hypothesis. Section 3 gives five reasons to think that the growth assumptions underlying the singularity hypothesis are too ambitious. Sections 4-5 consider recent philosophical arguments for the singularity hypothesis by Chalmers (2010) and Bostrom (2014), showing how these arguments fail to support the hypothesis' growth assumptions. Section 6 concludes by drawing out philosophical implications of this discussion for our understanding of consciousness, personal identity, digital minds, existential risk, and ethical longtermism.

2 Formulating the singularity hypothesis

The term 'singularity hypothesis' has come to be understood in a variety of ways. Following Ammon Eden and colleagues (2012), I will assume that a singularity hypothesis has three components. First, it specifies a quantity. Second, it claims that this quantity will experience accelerated growth. And third, it claims that accelerated growth will lead to a discontinuity in human history. Specifying each of these components in detail will identify the version of the singularity hypothesis that interests me.³

²The singularity hypothesis has also influenced discussions of existential risk from artificial intelligence at institutes such as the Future of Humanity Institute and the Center for the Governance of AI at Oxford, the Center for the Study of Existential Risk and Leverhulme Centre for the Future of Intelligence at Cambridge, the Center for Human-Compatible Artificial Intelligence and Existential Risk Initiative at Berkeley, the Center for AI Safety, Alignment Research Center, AI impacts, Anthropic, and Redwood Research. Note that the Singularity Institute has since been renamed to the Machine Intelligence Research Institute.

³In particular, it is important to separate this version of the singularity hypothesis from a variety of weaker economic claims that, while highly contentious, merit serious consideration. See Nordhaus (2021) for discussion.

2.1 Which quantity

The quantity that I am interested in is *intelligence*. More specifically, I am interested in the claim that the general intelligence of artificial agents will grow through processes of recursive self-improvement. This is perhaps the best-known form of the singularity hypothesis, introduced to philosophers by I.J. Good:

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 (Good 1966, p. 33).

Good’s claim is most naturally understood by assuming that there is an unambiguous concept of general intelligence and that this concept can be coherently applied to artificial agents. The claim is then that we should expect accelerating growth in the general intelligence of self-improving machines.

A number of commentators have raised doubts about the cogency of the concept of general intelligence (Nunn 2012; Prinz 2012), or the likelihood of artificial systems acquiring meaningful levels of general intelligence (Dreyfus 2012; Lucas 1964; Plotnitsky 2012). I have some sympathy for these worries.⁴ But these responses are controversial: many artificial intelligence researchers are happy to speak of general intelligence and express optimism about the likelihood of achieving human-level machine intelligence within this century.⁵ By contrast, I think that the falsity of the singularity hypothesis should be uncontroversial. Hence I am prepared to grant, for the sake of argument, both the cogency of the concept of general intelligence and the supposition that the general intelligence of artificial agents may meaningfully approach human intelligence. My skepticism will concentrate instead on the singularity hypothesis’ ambitious growth assumptions.

2.2 Accelerating growth

The singularity hypothesis posits a sustained period of accelerating growth in the general intelligence of artificial agents. This is a strong assumption, but the strength of the assumption turns on how growth rates are measured. Typically, growth rates refer to *relative growth*, the percentage change of a quantity over time, so that for example a change from 1 to 2 units of intelligence represents the same growth rate as a change from 2 to 4 units of intelligence. More rarely, growth rates refer to the weaker concept of *absolute growth*, the unscaled change in magnitude of a quantity over time, so that for example a

⁴However, it may be possible to reformulate arguments for the singularity hypothesis without appeal to the concept of general intelligence (Chalmers 2010).

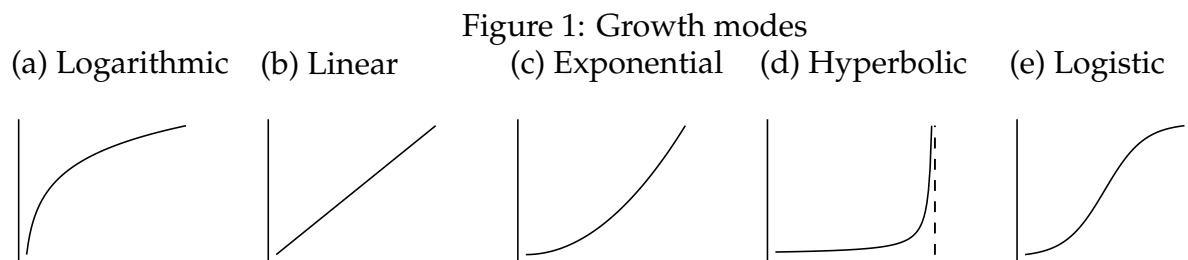
⁵For example, the median respondent to Grace et al (2016) gave a 50% chance of human-level machine intelligence within 40 years, and a 90% chance within 100 years. Similarly, the median respondent to Müller and colleagues’ (2016) survey of top-cited AI researchers gave a 50% chance of human-level intelligence by 2050 and a 90% chance by 2070. As always, there is some difficulty in interpreting this data, and in particular Grace and colleagues found moderate sensitivity to the manner in which the question is framed.

change from 1 to 2 units of intelligence represents the same growth rate as a change from 2 to 3 units of intelligence.

On both absolute and relative understandings, to say that machine intelligence grows at an accelerating rate is to rule out many familiar types of growth, such as logarithmic growth (Figure 1a) and linear growth (Figure 1b). But these understandings differ on what they say about exponential growth (Figure 1c). Sustained exponential growth is a very strong growth assumption, but even so, exponential growth only represents an accelerating growth rate on the weaker understanding that growth rates track absolute growth. And in fact, as fast as exponential growth may be, it is often not fast enough to produce the qualitative behavior demanded of an intelligence explosion. For example, if intelligence grew exponentially at a rate of 2 percent per year, then it would take almost 350 years for intelligence to grow three orders of magnitude. Even if intelligence doubled every two years, it would take twenty years for intelligence to grow three orders of magnitude. For this reason, although some authors model an intelligence explosion using exponential growth (Chalmers 2010), most authors have understood the intelligence explosion to involve still faster modes of growth.

The most common model of an intelligence explosion involves hyperbolic growth (Figure 1d), which approaches a vertical asymptote in finite time (Bostrom 2014; Solomonoff 1985). Hyperbolic growth does constitute accelerating growth on both relative and absolute understandings, and precisely for this reason it is difficult to overstate just how strong a growth assumption is made by hyperbolic models. For example, on Bostrom's (2014) model intelligence doubles in 7.5 months, grows a thousandfold within 17.9 months, and approaches infinity at 18 months. In some ways, these numbers may still be too conservative. Bostrom defines a 'fast' takeoff of artificial intelligence as one occurring within minutes, hours or days, and a 'moderate' takeoff as occurring in months or years. Bostrom argues that a fast or moderate takeoff is highly likely to occur. These are strong growth assumptions, and they will need a correspondingly strong argument to ground them.

It is important to note that the singularity hypothesis posits a *sustained* period of hyperbolic or exponential growth. This is important, because many quantities grow rapidly for some time, but then revert to calmer growth modes such as logistic growth (Figure 1e). If we project forward past trends of accelerating growth, we risk neglecting the likelihood that growth will flatten out (Modis 2012). For example, economic models suggest that world population grew exponentially or even hyperbolically for much of the twentieth century, rising from 1.6 billion in 1900 to 4 billion in 1975 (Figure 2a). One famous model from this period held that continued hyperbolic growth would lead to an infinite population by 2026 (Von Foerster et al. 1960). But it would have been a mistake to project continued hyperbolic population growth. World population is now projected to flatten



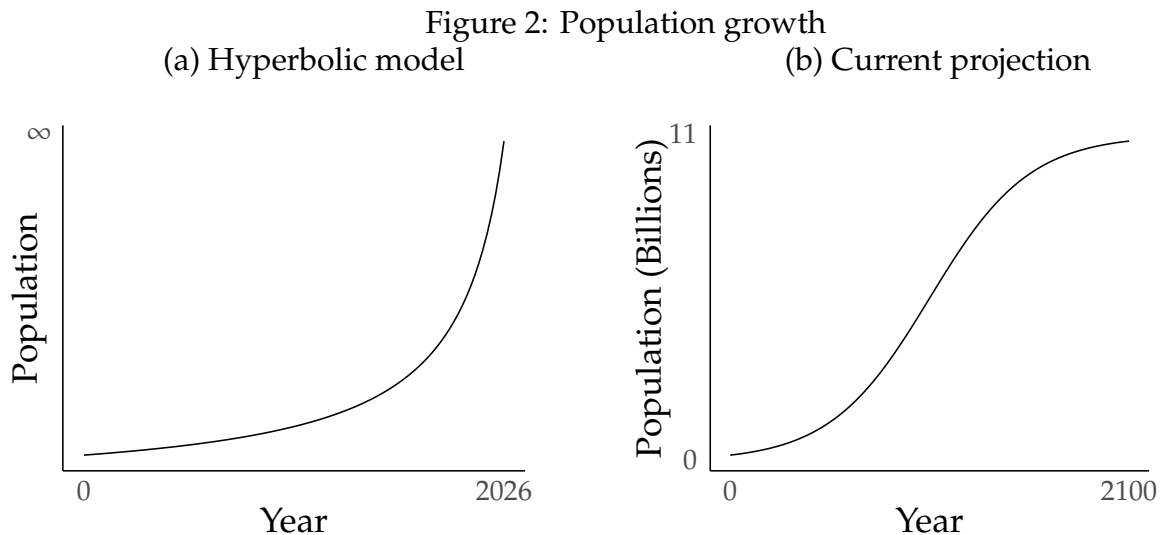
out at around 11 billion people by 2100 (Figure 2b) and perhaps even to decrease thereafter (United Nations 2019). And in the end, what began as a period of sharply accelerating growth will end with a relatively modest effect on total population. By contrast, the singularity hypothesis projects a period of sustained growth lasting much longer than recent population growth. Without sustained growth, an intelligence explosion would produce more modest gains.

2.3 Discontinuity

How long must this period of accelerated growth continue? The singularity hypothesis posits that growth will continue at least until a radical discontinuity in history is reached. It is difficult to overstate the magnitude of the envisaged transformation. If all goes well, it is held that artificial systems may become capable enough to upload copies of all living humans as full-fledged digital simulations of ourselves (Chalmers 2010). Humanity would then live on in digital form. And if all goes badly, artificial agents are predicted to become powerful enough to render humanity extinct, impotent or irrelevant at will (Bostrom 2014). We will discuss these implications in more detail in Section 6, but for now, just how much intelligence growth would be required to bring these developments about?

Opinions vary, but most advocates of the singularity hypothesis envision a machine several orders of magnitude more intelligent than the average human. For example, Richard Loosemore and Ben Goertzel take as their minimum target “a level of general intelligence 2-3 orders of magnitude greater than the human level” (Loosemore and Goertzel 2012, p. 86), and Chalmers considers a machine “at least as far beyond the most intelligent human as the most intelligent human is beyond a mouse” (Chalmers 2010, p. 11). The envisioned future is not one in which artificial agents grow to match humans, but one in which artificial agents leave unmodified humans far behind.

Summing up, the singularity hypothesis posits a period of sustained growth in the general intelligence of self-improving artificial agents. Intelligence growth will occur at an accelerating rate, yielding prolonged exponential or even hyperbolic growth. This



growth will continue until a fundamental discontinuity in human history is reached and human destiny is firmly in control of artificial agents orders of magnitude more intelligent than ourselves.

These are strong claims, and they should require a correspondingly strong argument to ground them. In Section 3, I give five reasons to be skeptical of the singularity hypothesis' growth claims. Then in Sections 4-5, I argue that existing defenses of the singularity hypothesis do little to overcome the case for skepticism. Section 6 concludes by drawing out philosophical implications of this discussion.

3 Growth

Existing responses to the singularity hypothesis often pursue one of two strategies. First, they push on the cogency of the concept of general intelligence or its applicability to the case at hand (Nunn 2012; Heylighen 2012; Prinz 2012). Second, they deny that artificial systems will reach human-level intelligence any time soon (Dreyfus 2012; Lucas 1964; Plotnitsky 2012). I want to explore a different source of resistance to the singularity hypothesis: denying its growth assumptions. In this section, I give five reasons to be skeptical of the rate of intelligence growth needed to ground the singularity hypothesis. Then in Sections 4-5, I argue that leading philosophical defenses of the singularity hypothesis do little to allay skepticism about growth rates.

3.1 Extraordinary claims require extraordinary evidence

The singularity hypothesis posits a sustained period of exponential or hyperbolic growth in the intelligence of artificial agents, continuing at least until machines exceed humans in intelligence by as much as humans exceed mice. These are extraordinary claims, and they should require correspondingly extraordinary evidence. It is not enough to show that sustained periods of hyperfast growth are physically possible, fail to be ruled out by our evidence, or are supported by a few suggestive considerations. To lend any significant credibility to their growth claims, advocates of the singularity hypothesis owe us many excellent reasons to suspect that, despite the unusual nature of the growth rates they posit, these growth rates do in fact describe the future growth of machine intelligence. Until this evidential burden is met, it is appropriate to place very low credence in the singularity hypothesis because of the extraordinary nature of its growth claims.

3.2 Good ideas become harder to find

Suppose you are fishing without replacement from a deep pond. At first, catching fish is easy. The fish you catch dwell in shallow waters and are easily fooled into biting. But as time goes on, the time between catches grows longer. Now the fish left in the pond are shrewder and dwell deeper underwater. Catching a fish goes from a matter of minutes to a matter of hours, days, or weeks.

Many social scientists think that beyond a point, nearly all idea-generating processes behave like fishing. As low-hanging fruit is plucked, good ideas become harder to find (Bloom et al. 2020; Kortum 1997; Gordon 2016). Research productivity, understood as the

amount of research input needed to produce a fixed output, falls with each subsequent discovery.

By way of illustration, the number of FDA-approved drugs per billion dollars of inflation-adjusted research expenditure decreased from over forty drugs per billion in the 1950s to less than one drug per billion in the 2000s (Scannell et al. 2012). And in the twenty years from 1971 to 1991, inflation-adjusted agricultural research expenditures in developed nations rose by over sixty percent, yet growth in crop yields per acre dropped by fifteen percent (Alston et al. 2000). The problem was not that researchers became lazy, poorly educated or overpaid. It was rather that good ideas became harder to find. Researchers in earlier decades identified many promising drugs and improvements to farming techniques, leaving later researchers to scrounge for more complex interventions. As good ideas became harder to find, research productivity declined.

Could the problem of improving artificial agents be an exception to the rule of diminishing research productivity? That is unlikely. Consider Moore's law, a historically observed doubling of hardware capacities every two years.⁶ A recent economic review singles out Moore's law as a prime example of declining research productivity: between 1971 and 2014, Moore's law was sustained across an eighteen-fold drop in semiconductor research productivity by astronomical increases in the amount of labor and capital devoted to semiconductor research (Bloom et al. 2020). If this situation seems unsustainable, that is because it is: many forecasters think that Moore's law will end in this decade, and most of the rest think it has already ended (Mack 2011; Shalf 2020; Theis and Wong 2017; Waldrop 2016). The declining pace of hardware growth is another piece of evidence for the relevance of declining research productivity to self-improving artificial intelligence.

Even if there is nothing special about the *problem* of improving artificial agents, could there be something special about the researchers we put this problem to? Perhaps diminishing research productivity afflicts human researchers, but not artificial agents. There is a sliver of truth to this objection: one cause of diminishing research productivity is the difficulty of maintaining large knowledge stocks (Jones 2009), a problem at which artificial agents excel. However, the underlying problem of fishing-out is a feature of problems, not agents, and as such it cannot be eliminated by enlisting artificial agents as researchers. After a while, any sensible investigatory process will have made more than its share of the easiest discoveries, and subsequent discoveries will become harder. Past that point, research productivity will diminish. Even many arguments for the singularity hypothesis concede this point, building diminishing research productivity directly into their models of intelligence growth (Davidson 2021; Roodman 2020).

Diminishing research productivity may not pose a severe threat to research processes that take us a hop, skip or a jump beyond current knowledge, but it does threaten research processes which seek to carry artificial intelligence many orders of magnitude beyond its current bounds. Even a relatively small rate of diminishing research productivity, compounded over many cycles of self-improvement, would become very large, exerting substantial downward pressure on the rate of intelligence growth.

⁶Moore's law is standardly operationalized by counting the number of transistors on a dense integrated circuit. However, there are other ways to operationalize Moore's law.

3.3 Bottlenecks

Growth processes often slow down because they hit bottlenecks: single obstacles which hinder further improvement until they can be overcome. Just one bottleneck is enough to slow growth.

One way of pressing the point is due to Philippe Aghion and colleagues (Aghion et al. 2019). Algorithms consist of various components such as search, computation and storage processes. These processes, in turn, break into subcomponents such as more elementary computations. It is an unalterable mathematical fact that an algorithm can run no more quickly than its slowest component. If nine-tenths of the component processes can be sped up, but the remaining processes cannot, then the algorithm can only be made ten times faster. This creates the opportunity for bottlenecks unless every single process can be sped up at once.

One reason to expect bottlenecks is that it is unclear how much room we have to improve certain processes. For example, computer scientists and mathematicians have been working on search algorithms for a very long time. While there is doubtless room for improvement, if our search algorithms are already in sight of the best-achievable performance then there is only so much we can do to make them faster. Another reason to expect bottlenecks is that many of the physical resource constraints to be discussed in Section 3.4 take time to overcome. Improving artificial systems requires generating and distributing large amounts of energy; mining key materials; building capital equipment; and constructing manufacturing plants. We may well be able to speed up these processes, but it is not plausible to expect them all to be made thousands of times more efficient overnight. And if even one process resists being sped up, the whole system of recursive self-improvement may be delayed.

To posit a fast and sustained process of recursive self-improvement requires the implausible assumption that all major bottlenecks to self-improvement may be overcome. If, as is likely, some bottlenecks remain, then the process of recursive self-improvement will slow.

3.4 Physical constraints

One reason why rapid growth processes in nature come to a halt is that they run up against physical constraints, such as finite resources and physical laws. To see the point in context, consider why most experts think that Moore's law is coming to an end. Why is it so unlikely that the number of transistors per circuit will continue to double biannually?

The first challenge involves resource constraints. Packing more transistors into a fixed area has led to massive increases in the amount of energy that must be fed through circuits during computation (Mack 2011). These energy demands are not only expensive, but also unsafe. There is only so much energy we can push through a circuit without overheating it given current technology. In addition to energy, we are also running out of capital (Waldrop 2016). The cost of new semiconductor plants has ballooned into the billions of dollars, and it is becoming increasingly difficult to continue meeting hardware targets by throwing more money at them.

In addition to resource constraints, we are also running up against basic laws of physics. For example, our best circuits use transistors whose diameter is only ten times that of a

typical atom (Yeap et al. 2019). As transistor size takes a nosedive towards the subatomic, it becomes increasingly hard to manufacture and interact with shrinking transistors, and previously irrelevant quantum uncertainty phenomena arise as potential limitations on the size of transistors.

The demise of Moore’s law holds an important lesson for the progress of self-improving machine intelligence. Any viable path to improving machine intelligence will eventually run up against resource constraints and laws of physics in ways that are not easily surmountable. When this happens, there is a good chance that the growth of intelligence will slow.

3.5 Sublinearity of intelligence growth from accessible improvements

One reason why the singularity hypothesis might seem plausible is that recent years have seen rapid improvements in quantities such as processing speed, memory and search depth. If intelligence grows proportionally to these quantities, then exponential growth in intelligence should require no more than a continuation of current trends. But this appeal masks an unfortunate fact: the relevant notion of intelligence grows sublinearly in underlying quantities such as processing speed that have been the targets of recent growth.

One way to see the point is to consider performance metrics that might plausibly correlate with intelligence. Here the story is strikingly consistent. Across almost any metric we care to measure, sustained exponential growth in computer capacities has led at best to linear growth in performance. For example, Neil Thompson and colleagues (2022) survey performance gains in computer game-play (Chess, Go) as well as three tasks that are highly dependent on sophisticated computation: protein folding, weather prediction, and modeling underground oil reservoirs. In each domain, Thompson and colleagues find exponential increases in the amount of compute power applied over time, but merely linear gains across metrics such as Chess and Go ELO rating, ability to solve protein folding problems, or to predict weather patterns and oil reserves. Across domains, the lesson seems to be that exponential increases in underlying computer capacities lead at best to linear improvements in intelligence-related tasks.

Natural objections to the argument from performance metrics do little to improve the situation. Perhaps advocates of the singularity hypothesis might object that all of the metrics in question correlate only imperfectly with intelligence. But in that case, they are welcome to supply alternative metrics which show consistent patterns of exponential increase. So far, no plausible metrics have been provided. Or perhaps it might be objected that the slow pace of performance increase is due to unrelated factors such as the increasing difficulty of successive improvements (Section 3.2). But even if that were true, it would merely relocate the problem, not solve it.

Another way to see that intelligence grows sublinearly in underlying capacities is to think about Moore’s law. In the past 50 years, the number of transistors in our best circuits increased from 3,500 in 1972 to 114 billion in 2022, and other hardware capacities showed similar growth. If we thought that intelligence grew linearly as a function of hardware capacity, we would have to say that computers have become 33 million times more intelligent over this period.

An immediate reaction to that claim is that it is implausible. Perhaps more carefully,

if advocates of the singularity hypothesis want to make such claims, they need to do two things. First, they need to clarify the relevant notion of intelligence on which it makes sense to speak of an intelligence increase on this scale having occurred. And second, they need to explain how the relevant notion of intelligence can do the work that their view demands. For example, they need to explain why we should expect increases in intelligence to lead to proportional increases in the ability to design intelligent agents (Section 4) and why we should attribute impressive and godlike powers to agents several orders of magnitude more intelligent than the average human (Section 6).

If, instead, the relevant notion of intelligence grows sublinearly in underlying computing capacities, then exponential growth in intelligence would require superexponential growth in underlying capacities. And hyperbolic growth? Don't ask.

3.6 Summarizing the case against the singularity hypothesis

The singularity hypothesis posits a sustained period of explosive growth in machine intelligence. In this section, we have met five reasons to be skeptical of the hypothesis' growth claims. First, the singularity hypothesis makes an extraordinary claim which should require correspondingly extraordinary evidence. Second, as time goes on ideas for further improvement will become harder to find. Third, even a single bottleneck to growth would be enough to halt an intelligence explosion. Fourth, a number of physical constraints including resource limitations will close off likely avenues for growth. And finally, rapid growth in machine intelligence might well require unfathomably fast growth in underlying quantities such as memory and computational speed.

Defenders of the singularity hypothesis owe us strong reasons to accept their growth claims in light of the obstacles that those claims have to overcome. In the next two sections, I survey leading philosophical defenses of the singularity hypothesis. I argue that these defenses are too weak to overcome skepticism about the singularity hypothesis' growth claims.

4 The observational argument

Chalmers argues for the singularity hypothesis by proposing the *proportionality thesis*: "increases in intelligence . . . always lead to proportionate increases in the ability to design intelligent systems" (Chalmers 2010, p. 21). If the proportionality thesis is true, then each successive AI system should represent at least as much of an improvement as its successor, leading to an intelligence explosion as long as these improvements can be brought about quickly enough.

Section 3 suggested that growth rates are likely to diminish, which in this context means that increases in artificial intelligence should eventually produce less-than-proportionate increases in the ability to design artificial systems. Chalmers must deny that growth rates will diminish, at least until many iterations of self-improvement have elapsed. Given the reasons for skepticism on this front, one might expect an extended argument against diminishing growth. But here is the entirety of Chalmers' argument against the possibility of diminishing growth:

If anything, 10% increases in intelligence-related capacities are likely to lead to all sorts of intellectual breakthroughs, leading to next-generation increases in intelligence that are significantly greater than 10%. Even among humans, relatively small differences in design capacities (say, the difference between Turing and an average human) seem to lead to large differences in the systems that are designed (say, the difference between a computer and nothing of importance). (Chalmers 2010, p. 27).

The first sentence is a restatement of the proportionality thesis, putting the burden of the argument on the second sentence. In that sentence, Chalmers uses a single observational data point, the difference between Turing and an average human, to defend the proportionality thesis across the board. Call this the *observational argument*.

The observational argument faces two challenges. First, it is local rather than global. It points to a single moment in the history of machine intelligence and argues that, at this local point, increases in the intelligence of machine designers led to a greater-than-proportional increase in the intelligence of resulting systems. The problem with local arguments is that they cannot demonstrate the possibility of sustained rates of high growth, because they only sample growth rates at a single point on a curve. Skeptics of the singularity hypothesis do not deny that there was *any* point in the history of machine intelligence at which the proportionality thesis held. What we deny is that the proportionality thesis will continue to hold through long processes of recursive self-improvement. A single local observation about growth rates does not constitute evidence that these rates will continue indefinitely (Dreyfus 2012).

Moreover, the local point which Chalmers sampled has been chosen to maximally downplay the possibility of diminishing growth rates. By considering growth rates in the very early history of computing, Chalmers samples a point before low-hanging fruit is likely to have been used up, resources depleted, or bottlenecks encountered, avoiding many of the best arguments for diminishing growth. We can readily grant that early-stage research often exhibits constant or increasing growth while retaining substantial confidence that growth rates will later fall.

A second worry for the observational argument is that it equivocates between intelligence and design capacities. The datum cited is that a small difference in *design capacities* between Turing and an average human led to large differences in the *intelligence* of the system designed. This is not an innocent slip, for we would not want to deny that many of Turing's contemporaries were more intelligent than Turing, but much less capable of designing intelligent systems. Therefore, it is not clear that the right thing to conclude from this discussion is that increases in intelligence lead to proportionate increases in the capacity to design intelligent systems. Otherwise, we would have to explain why many people more intelligent than Turing lacked Turing's capacity to design such systems.

Now as Chalmers notes, he could do equally well if all mentions of intelligence were replaced by some other term. So it would be enough to take the example as one in which a small difference in design capacities between Turing and an average human led to large differences in the *design capacities* of the system designed. But this inference fails on both ends. It is far from obvious that the systems designed by Turing had any design capacities at all. Nor is it obvious that the difference in design capacities between Turing and an average human are aptly described as small. Indeed, one might rightly take Turing's

design efforts as evidence that his design capacities far outstripped those of an average contemporary.

In this section, we considered an argument for the singularity hypothesis based on the proportionality thesis that increases in intelligence always lead to proportionate increases in the ability to design intelligent systems. We considered Chalmers' observational argument for the proportionality thesis, which cites Turing's early progress as a historical example in which the proportionality thesis held. However, we saw that this argument fails on two fronts: it is local rather than global, and it equivocates between intelligence and design capacities. Hence Chalmers' observational argument does not vindicate the singularity hypothesis.

5 Recalcitrance and optimization power

Bostrom (2014) argues that the rate of change of artificial intelligence will be driven by two quantities. The first is the amount of *optimization power*, or quality-weighted design effort applied to improving artificial systems. The second quantity is the *recalcitrance* of intelligence growth, the amount of optimization power that must be applied to produce a unit change in intelligence at the current margin. With this understanding of optimization power and recalcitrance, intelligence grows fractionally in optimization power and recalcitrance:

$$\text{Rate of intelligence change} = \frac{\text{Optimization power}}{\text{Recalcitrance}}.$$

If a high amount of optimization power is applied throughout a sustained period of low recalcitrance, the result will be an intelligence explosion.

A general difficulty in theorizing about the singularity hypothesis is that many arguments for the singularity hypothesis are thin on empirical details about how the intelligence explosion is meant to come about. Bostrom provides a refreshing contrast by discussing a number of specific reasons to expect low recalcitrance and high optimization power to obtain. But the devil is in the details, and in this section I argue that Bostrom's empirical case for low recalcitrance and high optimization power is insufficient.⁷

More specifically, I argue that the details offered by Bostrom break into three categories. The first and largest class consists of mundane but plausible descriptions of how the future might develop which have been over-interpreted to support the possibility of intelligence explosion. The second class consists of thinly detailed restatements of the core hope behind the singularity hypothesis. And the third class consists of factually mistaken interpretations of past historical trends. I conclude that despite the rich detail in which his argument is made, Bostrom fails to provide adequate support for a sustained period of high optimization power combined with low recalcitrance.

5.1 Plausible but over-interpreted scenarios

Bostrom offers five reasons to think that recalcitrance will remain low throughout a period of sustained intelligence growth. Three of these suggestions are within the realm of

⁷To the best of my knowledge, this section surveys *every* detailed suggestion from Chapter 4 of *Superintelligence* in support of low recalcitrance and high optimization power.

plausibility, but do not lend significant support to the possibility of intelligence explosion. First, suppose that artificial intelligence is reached through whole-brain emulation, so that the first human-level artificial intelligence is a computer emulation of a single human brain.⁸ En route to the first emulation, we may experience high recalcitrance as we struggle to adjust software and to meet hardware demands. But once the first emulation is produced, recalcitrance may drop as we make a series of rapid improvements. For example, we might simplify the algorithm by coarse-graining neural representations, allowing it to run more quickly. Or we might emulate the brains of distinct individuals, increasing the stock of artificial agents we can draw on.

This is our first example of a detailed and not-entirely-implausible scenario that has been overinterpreted to support the possibility of a sustained period of low recalcitrance. There is no doubt that whole-brain emulations, like any new technology, can be improved once they are introduced. Nor is there any doubt that the first improvements may proceed rapidly by exploiting low-hanging fruit. But we have not been given any reason to think that these improvements would be rapid, powerful and sustained enough to bring about a powerful superintelligence. Indeed, Bostrom himself concedes that improvements to emulation technologies may soon hit a period of diminishing returns.

A second example of how artificial intelligence may be improved is through what Bostrom terms *content* improvements, improvements to software that go beyond changes to core algorithms. Bostrom offers the example of increased knowledge stocks: an agent trained on a small database may be improved by training it on large data sets, such as the entire Library of Congress. Here again we have a detailed and relatively plausible scenario for improvement, but one with limited implications for the growth of artificial intelligence. Bostrom is quite right that there is a trend towards training agents on larger datasets, and that improving datasets is a good tool for improving the capabilities of artificial agents. But as impactful as such improvements are, we are given no reason to suspect that absorbing even the entire Library of Congress would result in the sudden emergence of superintelligence. Indeed, there is some reason for doubt on this score. After all, the Library of Congress is already part of the collective knowledge stock of humanity, but this knowledge has only taken us so far.

A third example of how artificial intelligence may be improved is by improving hardware: we may run artificial agents on faster, more powerful or more numerous hardware systems. Here we encounter a third and final example of something that is likely to pass, but unlikely to ground an intelligence explosion. We saw in Section 3 that artificial systems have already experienced a sustained period of rapid hardware growth, but that this growth alone did not produce astronomical gains, for example because intelligence grows sublinearly in hardware capacities. We also saw that many experts expect the pace of hardware growth to slow rather than accelerate, so an appeal to growing hardware capacities is unlikely to provide convincing evidence for an increase in the speed of intelligence growth.

⁸For skepticism about this route to superintelligence, see Mandelbaum (2022).

5.2 Restating the core hope

Bostrom offers two further reasons to think that recalcitrance will remain low during a period of sustained intelligence growth. First, Bostrom suggests that artificial intelligence may be produced in a single leap by a clever software insight. If we are only one leap of programming away from superintelligence, then a single programmer in a basement may one day initiate and complete the intelligence explosion. And second, Bostrom suggests, an intelligence explosion may result from a shift in modes of processing. It may be that right now, artificial agents progress primarily by improving their domain-general reasoning capacities, and that as the domain-general reasoning capacities of artificial agents begin to predominate, artificial agents will be able to turn those domain-general reasoning capacities towards the problem of their own improvement, resulting in rapid intelligence growth.

Both of these suggestions express optimistic hopes that, without further evidence, go little beyond the core claim of the singularity hypothesis. It is of course possible that a single software insight could produce superintelligence. It is also possible that artificial agents, once they reach a certain level of domain-general reasoning ability, could improve themselves through one or a series of rapid software insights. But to take these claims from bare possibility to scientific plausibility, we need to provide evidence for thinking that they are likely to come to pass. And so far, Bostrom has not provided any evidence for either claim.

5.3 Mis-interpreting history

How low would the recalcitrance of intelligence improvement have to be in order to generate an intelligence explosion? Bostrom asks us to suppose that optimization power increases linearly in the intelligence of an artificial systems, so that a system which doubles in intelligence doubles in optimization power. This seems at first like a more plausible version of Chalmers' proportionality thesis, since it removes the assumption of constant recalcitrance. However, Bostrom notes that with these assumptions in place, even constant recalcitrance combined with a linear relationship between intelligence and optimization power would produce only exponential intelligence growth, and under many assumptions the rate of exponential growth might be quite modest.

To generate hyperbolic growth, Bostrom makes a stronger assumption than Chalmers: that recalcitrance varies inversely with the level I of an agent's current intelligence, so that recalcitrance is equal to $1/I$. On this model, we do indeed get an intelligence explosion: the more intelligent an agent becomes, the easier it gets to make further progress, leading to hyperbolic growth in machine intelligence.

However, the assumption of rapidly decreasing recalcitrance is a surprising claim in need of justification. Here, Bostrom appeals to Moore's law:

Suppose that the optimization power applied to the system is roughly constant ... prior to the system becoming capable of contributing substantially to its own design, and that this leads to the system doubling in capacity every 18 months. (This would be roughly in line with historical improvement rates from Moore's law combined with software advances.) This rate of improve-

ment, if achieved by means of roughly constant optimization power, entails recalcitrance declining as the inverse of system power. (Bostrom 2014, p. 76).

In this passage, Bostrom asks us to imagine two things: that the intelligence of artificial agents has been doubling every eighteen months, and that this doubling has taken place without an increase in the optimization power applied to improving them. These assumptions would entail that historical rates of recalcitrance have been decreasing inversely in the intelligence of artificial agents, and if we project the trend forwards in time the argument for hyperbolic growth goes through.

However, each of Bostrom's assumptions mis-interprets historical data. Regarding the doubling time of intelligence, we saw in Section 3 that intelligence grows sublinearly in hardware capacities, so that for example from the fact that transistor counts grew thirty-three-millionfold over fifty years we cannot conclude that the intelligence of artificial agents grew millions-fold over the same period. We also saw in Section 3 that historical increases in hardware capacities were not achieved by a constant exertion of optimization power. Leading economic models suggest an eighteen-fold increase in the quality-weighted research effort needed to produce a doubling of hardware capacities over this period (Bloom et al. 2020). Hence the historical evidence does not support the assumptions of rapidly-doubling intelligence against constant optimization power that are needed to ground an intelligence explosion. Rather, the evidence suggests increasing recalcitrance of hardware gains, and thus strongly increasing recalcitrance of intelligence gains, since intelligence does not grow in lockstep with hardware. On Bostrom's model, these assumptions would not even ground exponential intelligence growth.

In this section, we considered Bostrom's argument for an intelligence explosion. This argument claimed that increasing amounts of optimization power will be applied during a period of constant or falling recalcitrance, leading to rapid growth in artificial intelligence. We saw that the considerations marshaled in support of Bostrom's conclusion fall into three categories. The first category consists of detailed descriptions of moderately plausible future scenarios, which have been wrongly interpreted as supporting an intelligence explosion. The second category consists of undetailed future scenarios that amount to thinly veiled assertions of the core hope behind the singularity hypothesis. And the final category consists of a mis-interpretation of historical growth trends which, once corrected, would tell against rather than in favor of intelligence explosion. Taken together, these considerations do not lend substantial support to the singularity hypothesis.

6 Philosophical implications

The singularity hypothesis posits a sustained period of accelerating growth in the general intelligence of artificial agents brought about by recursive processes of self-improvement. This growth is hypothesized to continue until artificial agents have become orders of magnitude more intelligent than their human designers, leading to a fundamental discontinuity in human history.

We saw in Section 3 that there are good reasons to be skeptical of the singularity hypothesis. The singularity hypothesis makes an extraordinary claim which should require correspondingly extraordinary evidence. As time goes on ideas for further improvement

will become harder to find. Like most growth processes, the growth of artificial intelligence is likely to be stalled by bottlenecks. Over time, physical limitations such as finite resources and the laws of physics will constrain the pace of improvement. And rapid growth in machine intelligence might well require unfathomably fast growth in underlying quantities such as memory and computational speed.

Sections 4-5 surveyed two leading philosophical arguments for the singularity hypothesis: Chalmers' observational argument and Bostrom's argument for a sustained period of low recalcitrance combined with high optimization power. We saw that each argument falls short of vindicating the singularity hypothesis. If that is right, then it would be inappropriate at this time to place substantial confidence in the singularity hypothesis.

This discussion has at least three classes of important philosophical implications that are worth drawing out.

6.1 Uploading: consciousness, identity, and value

If all goes well, an intelligence explosion raises the possibility of mental uploading whereby human minds are uploaded as digital simulations within a powerful computer (Chalmers 2010; Mandelbaum 2022; Pigliucci 2014). Indeed, it is largely this prospect that drew Chalmers to discuss the singularity hypothesis. The chance of mental uploading raises a host of questions. Would the resulting upload be conscious (Chalmers 2010; Dennett 2012)? Would it be me (Cerullo 2015; Goldwater 2021)? Could mental uploading allow me to become very long-lived (Smart 2021), and if so would that be a good thing (Scheffler and Kolodny 2013)?

The singularity hypothesis transforms these questions into pressing practical matters. If the singularity hypothesis is on the right track, then mental uploading is a choice that some of us may face within our own lifetimes. Making the right choice could produce millennia of bliss, but making the wrong choice could lead to our own deaths. However, I have argued that the singularity hypothesis is unlikely to be true. That raises the possibility of postponing judgment on some of the most difficult questions about uploading, identity and value until we have a better understanding of the scientific possibility of uploading, the way in which uploading would be implemented, and basic philosophical issues surrounding digital minds.

6.2 Existential risk from artificial intelligence

If all does not go well, an intelligence explosion could threaten humanity's future. It is increasingly argued that progress in artificial intelligence presents an *existential risk* to humanity (Bostrom 2014; Greaves and MacAskill 2021; Ord 2020), threatening "the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development" (Bostrom 2013, p. 15).

The most famous argument that artificial intelligence poses an existential risk to humanity is due to Nick Bostrom (2012; 2014) and Eliezer Yudkowsky (2013). The Bostrom-Yudkowsky argument has three premises. The first premise is the singularity hypothesis that self-improving artificial agents will rapidly explode in intelligence. The singularity hypothesis is taken to show that after an intelligence explosion, what happens next depends largely on the motivations of superintelligent artificial agents. Because these agents

will be orders of magnitude more intelligent than humans, humanity will be largely powerless to resist them.

The next premise of the Bostrom-Yudkowsky argument holds that in principle, super-intelligent agents could have any motivation at all (Bostrom 2012; Müller and Cannon 2021).

(Orthogonality Thesis) Intelligence and final goals are orthogonal axes along which possible agents can freely vary. In other words, more or less any level of intelligence could in principle be combined with more or less any final goal. (Bostrom 2012, p. 73).

The Orthogonality Thesis is taken to imply that there is a live threat of producing artificial agents with the wrong motivations. But just how live is this threat?

The final premise of the Bostrom-Yudkowsky argument holds that a wide range of motivations could be bad ones. Almost any agent with goals different from our own would be driven to seek instrumentally valuable items such as power and resources to put itself in a better position to achieve its goals (Bostrom 2012; Omohundro 2008).

(Instrumental convergence) Several instrumental values [such as power and resources] can be identified which are convergent in the sense that their attainment would increase the chances of the agent's goal being realized for a wide range of final goals and a wide range of situations, implying that these instrumental values are likely to be pursued by many intelligent agents. (Bostrom 2012, p. 76).

Instrumental convergence is taken to show the importance of ensuring that superintelligent agents have a high degree of respect for human well-being, or else hold values closely aligned to our own. Otherwise, they would be driven to take power and resources from humanity in order to achieve their goals, impoverishing or exterminating humanity in the process.

There is a great deal to be said about the Bostrom-Yudkowsky argument (Müller and Cannon 2021; Vold and Harris forthcoming). But a surprisingly underexplored route of resisting the argument is to cut the argument off at its source. By denying the singularity hypothesis, we deprive the Bostrom-Yudkowsky argument of the original motivation for taking value-misaligned artificial agents to pose an existential threat to humanity. This would not imply that there is no interest in thinking about moral reasoning by artificial agents. But it does help to allay fears of existential risk from artificial agents, at least insofar as those fears are driven by the Bostrom-Yudkowsky argument.

6.3 Longtermism

There is growing philosophical (Greaves et al. forthcoming; Greaves and MacAskill 2021; Mogensen 2021) and public (Cargill and John 2021; MacAskill 2022; Ord 2020) interest in ethical *longtermism*. As an axiological matter, longtermism holds roughly that the best thing we can do is often what is best, or near-best for the very long-term future of humanity.⁹ And as a deontic matter, longtermism holds that we should often take actions

⁹See Greaves and MacAskill (2021) for a full definition.

which are best, or near-best for the long-term future of humanity. Roughly, the idea is that because the future of humanity is very large, anything we can do to improve that future will have a great deal of value. And because many deontic theories give some weight to the promotion of future well-being, opportunities to promote significant amounts of future well-being can take precedence over competing duties in determining how we ought to act today.

Many recent arguments against longtermism have questioned decision-theoretic (Steele forthcoming; Pettigrew 2022), consequentialist (Heikkinen 2022; Mogensen 2021; Unruh forthcoming) or population-ethical (Lederman and Frick forthcoming) assumptions used to motivate it. What is less often explored is the possibility of empirically-driven resistance to longtermism.

When longtermists are pressed to deliver concrete action recommendations, they often suggest acting to mitigate existential risk (Greaves and MacAskill 2021; Ord 2020). Existential risks posed by artificial intelligence are often held to be among the most pressing. For example, Toby Ord (2020) estimates a ten percent risk of existential catastrophe caused by artificial agents by 2100. Will MacAskill (2022) puts the risk at three percent. And participants at the Oxford Global Catastrophic Risk Conference estimated the chance of outright extinction caused by artificial agents before 2100 at five percent (Sandberg and Bostrom 2008). The discussion in this paper suggests that these estimates do not rest on a firm scientific basis, at least not insofar as they rely on considerations similar to the Bostrom-Yudkowsky argument.

It may be worth generalizing the strategy in this paper to explore the possibility of empirically-driven challenges to longtermism. These challenges would aim to undercut the scientific basis for existential risk mitigation and other favored longtermist interventions by showing that the empirical case for longtermist interventions often lags substantially behind leading longtermist estimates. A family of such challenges might combine to put pressure on the plausibility of longtermism.

6.4 Future directions

So far, we have seen that denying the singularity hypothesis has important implications for philosophical discussions of consciousness, uploading, personal identity, existential risk and ethical longtermism. It might be productive for future work to explore further implications of the singularity hypothesis. For example, what can reflection on the singularity hypothesis teach us about the nature and possibility of artificial intelligence (Dreyfus 2012)? What are the most plausible routes through which artificial general intelligence might eventually be reached (Mandelbaum 2022)? And how is the singularity hypothesis connected to more traditional ethical questions raised by growth in artificial intelligence? But as for the singularity hypothesis itself, I hope to have shown that the growth assumptions underlying the hypothesis are substantially less plausible than its advocates suppose.

References

- Aghion, Philippe, Jones, Benjamin, and Jones, Charles. 2019. "Artificial intelligence and economic growth." In Ajay Agrawal, Joshua Gans, and Avi Goldfarb (eds.), *The economics of artificial intelligence: An agenda*. University of Chicago Press.
- Alston, Julian, Chan-Kang, Connie, Marra, Michele, Pardey, Philip, and Wyatt, TJ. 2000. *A meta-analysis of rates of return to agricultural R&D*. International Food Policy Research Institute.
- Armstrong, Stuart, Bostrom, Nick, and Shulman, Carl. 2016. "Racing to the precipice: A model of artificial intelligence development." *AI and Society* 31:201–6.
- Awret, Uziel. 2012. "Introduction." *Journal of Consciousness Studies* 19:7–15.
- Baum, Seth, Goertzel, Ben, and Goertzel, Ted. 2011. "How long until human-level AI? Results from an expert assessment." *Technological Forecasting and Social Change* 78:185–95.
- Bellamy, Rachel et al. 2019. "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias." *IBM Journal of Research and Development* 63:4:1–4:15.
- Bloom, Nicholas, Jones, Charles, Van Reenen, John, and Webb, Michael. 2020. "Are ideas getting harder to find?" *American Economic Review* 110:1104–44.
- Bostrom, Nick. 2012. "The superintelligent will: Motivation and instrumental raitonality in advanced artificial agents." *Minds and Machines* 22:71–85.
- . 2013. "Existential risk prevention as a global priority." *Global Policy* 4:15–31.
- . 2014. *Superintelligence*. Oxford University Press.
- Callaghan, Victor, Miller, James, Yampolskiy, Roman, and Armstrong, Stuart (eds.). 2017. *The technological singularity: Managing the journey*. Springer.
- Cargill, Natalie and John, Tyler (eds.). 2021. *The long view: Essays on policy, philanthropy and the long-term future*. FIRST.
- Cerullo, Michael. 2015. "Uploading and branching identity." *Minds and Machines* 25:17–36.
- Chalmers, David. 2010. "The singularity: A philosophical analysis." *Journal of Consciousness Studies* 17:7–65.
- . 2012. "The singularity: A reply to commentators." *Journal of Consciousness Studies* 7-8:141–67.
- Davidson, Tom. 2021. "Could advanced AI drive explosive economic growth?" Technical report, Open Philanthropy, <https://www.openphilanthropy.org/research/could-advanced-ai-drive-explosive-economic-growth/>.
- Dennett, Daniel. 2012. "The mystery of David Chalmers." *Journal of Consciousness Studies* 19:86–95.

- Dreyfus, Hubert. 2012. "A history of first step fallacies." *Minds and Machines* 22:87–99.
- Eden, Amnon, Moor, James, Søraker, Johnny, and Steinhart, Eric (eds.). 2012. *Singularity hypotheses: A scientific and philosophical assessment*. Springer.
- Goldwater, Johan. 2021. "Uploads, faxes, and you: Can personal identity be transmitted?" *American Philosophical Quarterly* 58:233–50.
- Good, I.J. 1966. "Speculations concerning the first ultraintelligent machine." *Advances in Computers* 6:31–88.
- Gordon, Robert J. 2016. *The rise and fall of American growth*. Princeton University Press.
- Grace, Katja, Salvatier, John, Zhang, Baobao, and Evans, Owain. 2016. "2016 Expert survey on progress in AI." [Aiimpacts.org/2016-expert-survey-on-progress-in-ai](https://aiimpacts.org/2016-expert-survey-on-progress-in-ai).
- Greaves, Hilary and MacAskill, William. 2021. "The case for strong longtermism." Global Priorities Institute Working Paper 5-2021, <https://globalprioritiesinstitute.org/hilary-greaves-william-macaskill-the-case-for-strong-longtermism-2/>.
- Greaves, Hilary, Thorstad, David, and Barrett, Jacob (eds.). forthcoming. *Longtermism*. Oxford University Press.
- Heikkinen, Karri. 2022. "Strong longtermism and the challenge from anti-aggregative moral views." Global Priorities Institute Working Paper 5-2022, <https://globalprioritiesinstitute.org/karri-heikkinen-strong-longtermism-and-the-challenge-from-anti-aggregative-moral-views/>.
- Heylighen, Francis. 2012. "A brain in a vat cannot break out: Why the singularity must be extended, embedded and embodied." *Journal of Consciousness Studies* 17:126–42.
- Horvitz, Eric and Selman, Bart. 2012. "Interim report from the panel chairs: AAAI presidential panel on long-term AI futures." In Amnon Eden, James Moor, Johnny Søraker, and Eric Steinhart (eds.), *Singularity hypotheses: A scientific and philosophical assessment*, 301–6. Springer.
- Jones, Benjamin. 2009. "The burden of knowledge and the 'Death of the renaissance man': Is innovation getting harder?" *Review of Economic Studies* 76:283–317.
- Korinek, Anton and Stiglitz, Joseph. 2019. "Artificial intelligence and its implications for income distribution and unemployment." In Ajay Agrawal, Joshua Gans, and Avi Goldfarb (eds.), *The economics of artificial intelligence: An agenda*, 349–90. University of Chicago Press.
- Kortum, Samuel. 1997. "Research, patenting, and technological change." *Econometrica* 65:1389–1419.
- Kurzweil, Ray. 2005. *The singularity is near: When humans transcend biology*. Viking.
- Lederman, Harvey and Frick, Johann. forthcoming. "Deontic longtermism and the non-identity problem." In Jacob Barrett, Hilary Greaves, and David Thorstad (eds.), *Longtermism*. Oxford University Press.

- Lipton, Zachary. 2018. "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue* 16:31–57.
- Loosemore, Richard and Goertzel, Ben. 2012. "Why an intelligence explosion is probable." In Amnon Eden, James Moor, Johnny Søraker, and Eric Steinhart (eds.), *Singularity hypotheses: A scientific and philosophical assessment*, 83–96. Springer.
- Lucas, J.R. 1964. "Minds, machines and Gödel." In A.R. Anderson (ed.), *Minds and Machines*, 43–59. Prentice Hall.
- MacAskill, William. 2022. *What we owe the future*. Basic books.
- Mack, Chris. 2011. "Fifty years of Moore's Law." *IEEE Transactions on Systems Science and Cybernetics* 24:202–7.
- Mandelbaum, Eric. 2022. "Everything and more: The prospects of whole brain emulation." *Journal of Philosophy* 119:444–59.
- Modis, Theodore. 2012. "Why the singularity cannot happen." In Amnon Eden, James Moor, Johnny Søraker, and Eric Steinhart (eds.), *Singularity hypotheses: A scientific and philosophical assessment*, 311–40. Springer.
- Mogensen, Andreas. 2021. "Moral demands and the far future." *Philosophy and Phenomenological Research* 103:567–85.
- Müller, Vincent and Bostrom, Nick. 2016. "Future progress in artificial intelligence: A survey of expert opinion." In Vincent Müller (ed.), *Fundamental issues of artificial intelligence*, 555–72. Springer.
- Müller, Vincent and Cannon, Michael. 2021. "Existential risk from AI and orthogonality: Can we have it both ways?" *Ratio* 35:25–36.
- Nordhaus, William. 2021. "Are we approaching an economic singularity? Information technology and the future of economic growth." *American Economic Journal: Macroeconomics* 13:299–332.
- Nunn, Chris. 2012. "More splodge than singularity?" *Journal of Consciousness Studies* 19:57–60.
- Omohundro, Stephen. 2008. "The basic AI drives." In Pei Wang, Ben Goertzel, and Stan Franklin (eds.), *Proceedings of the 2008 conference on artificial intelligence*, 483–92. IOS Press.
- Ord, Toby. 2020. *The precipice*. Bloomsbury.
- Pettigrew, Richard. 2022. "Effective altruism, risk, and human extinction." Global Priorities Institute Working Paper 2-2022, <https://globalprioritiesinstitute.org/effective-altruism-risk-and-human-extinction-richard-pettigrew-university-of-bristol/>.
- Pigliucci, Massimo. 2014. "Mind uploading: A philosophical counter-analysis." In Russell Blackford and Damien Broderick (eds.), *Intelligence unbound: The future of uploaded and machine minds*, 119–30. Wiley.

- Plotnitsky, Arkady. 2012. "The singularity wager: A response to David Chalmers." *Journal of Consciousness Studies* 7-8:61–76.
- Prinz, Jesse. 2012. "Singularity and inevitable doom." *Journal of Consciousness Studies* 19:77–86.
- Roodman, David. 2020. "On the probability distribution of long-term changes in the growth rate of the global economy: An outside view." Technical report, Open Philanthropy, <https://www.openphilanthropy.org/sites/default/files/Modeling-the-human-trajectory.pdf>.
- Russell, Stewart. 2019. *Human compatible: Artificial intelligence and the problem of control*. Viking.
- Sandberg, Anders and Bostrom, Nick. 2008. "Global catastrophic risks survey." Technical Report 2008-1, Future of Humanity Institute, <https://www.global-catastrophic-risks.com/docs/2008-1.pdf>.
- Scannell, Jack, Blanckley, Alex, Boldon, Helen, and Warrington, Brian. 2012. "Diagnosing the decline in pharmaceutical R&D efficiency." *Nature Reviews Drug Discovery* 11:191–200.
- Scheffler, Samuel and Kolodny, Niko (eds.). 2013. *Death and the afterlife*. Oxford University Press.
- Shalf, John. 2020. "The future of computing beyond Moore's law." *Philosophical Transactions of the Royal Society A* 378:20190061.
- Smart, Paul. 2021. "Predicting me: The route to digital immortality?" In Inês Hipólito, Robert William Clowes, and Klaus Gärtner (eds.), *The mind-technology problem: Investigating minds, selves and 21st century artefacts*, 185–207. Springer.
- Solomonoff, Ray. 1985. "The time scale of artificial intelligence: Reflections on social effects." *Human Systems Management* 5:149–53.
- Steele, Katie. forthcoming. "Risk aversion, the long term and agent-centred prerogatives." In Jacob Barrett, Hilary Greaves, and David Thorstad (eds.), *Longtermism*. Oxford University Press.
- Stein-Perlman, Zach, Weinstein-Raun, Benjamin, and Grace, Katja. 2022. "2022 Expert Survey on Progress in AI." *AI Impacts*, 3 Aug 2022, <https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/>.
- Theis, Thomas and Wong, H.S. Philip. 2017. "The end of Moore's Law: A new beginning for information technology." *Computing in Science and Engineering* 19:41–50.
- Thompson, Neil, Ge, Shuning, and Manso, Gabriel. 2022. "The importance of (exponentially more) computing power." ArXiv Pre-print, DOI 2206.14007.
- United Nations, Population Division, Department of Economic and Social Affairs. 2019. "World Population Prospects 2019."

Unruh, Charlotte. forthcoming. "Deontology, harm, and generational sovereignty." In Jacob Barrett, Hilary Greaves, and David Thorstad (eds.), *Longtermism*. Oxford University Press.

Vinge, Vernor. 1993. "The coming technological singularity: How to survive in the post-human era." *Whole Earth Review* 81:88–95.

Vold, Karina and Harris, Daniel. forthcoming. "How does artificial intelligence pose an existential risk?" In *Oxford Handbook of Digital Ethics*, forthcoming. Oxford University Press.

Von Foerster, Heinz, Mora, Patricia, and Amiot, Lawrence. 1960. "Doomsday: Friday, 13 November, A.D. 2026." *Science* 132:1291–5.

Waldrop, M. Mitchell. 2016. "The chips are down for Moore's law." *Nature* 530:144.

Yampolskiy, Roman. 2016. *Artificial superintelligence: A futuristic approach*. Chapman and Hall.

Yeap, Goffrey et al. 2019. "5nm CMOS Production Technology Platform featuring full-fledged EUV, and High Mobility Channel FinFETs with densest $0.021\mu\text{m}^2$ SRAM cells for Mobile SoC and High Performance Computing Applications." *2019 IEEE International Electron Devices Meeting (IEDM)* 36.7.1–36.7.4.

Yudkowsky, Eliezer. 2013. "Intelligence explosion microeconomics." Technical Report 2013-1, Machine Intelligence Research Institute, <http://intelligence.org/files/IEM.pdf>.

Zhang, Baobao, Dreksler, Noemi, Anderljung, Markus, Kahn, Lauren, Giattino, Charlie, Dafoe, Allan, and Horowitz, Michael C. 2022. "Forecasting AI Progress: Evidence from a Survey of Machine Learning Researchers." <https://doi.org/10.48550/arXiv.2206.04132>.