Convergence Analysis

Pathways to short TAI timelines

AUTHOR

ZERSHAANEH QURESHI, CONVERGENCE ANALYSIS

PUBLISHED

FEBRUARY 2025

Content finalised January 2025

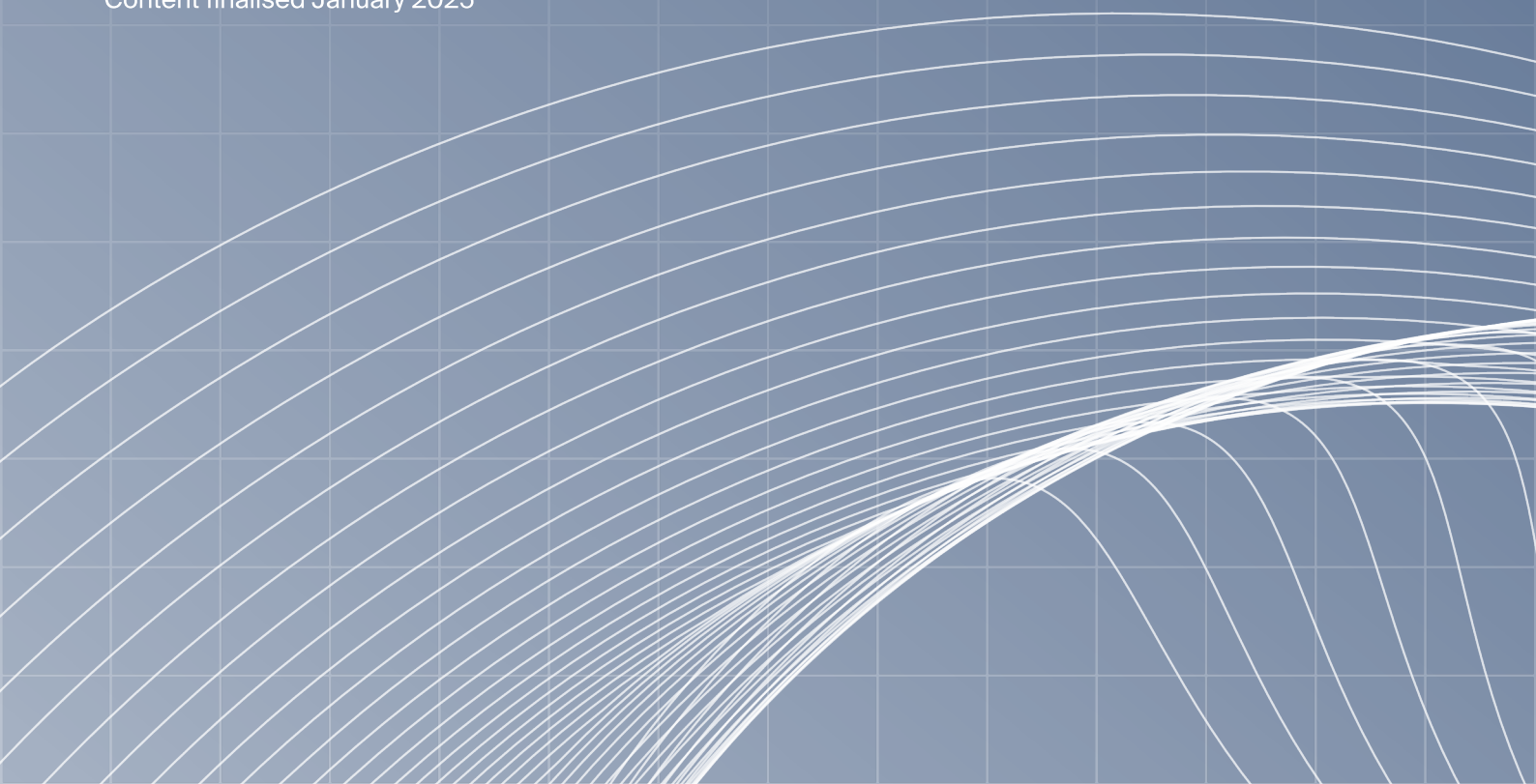


Table of Contents

Purpose of this report

The timeline for the arrival of advanced forms of AI such as 'transformative AI' (TAI) has been the subject of significant debate for decades. This report explores the prospect of what I call 'short timelines' to TAI, referring to the development of TAI within the next ten years (i.e., by 2035). Over several chapters, it gathers evidence, charts out the territory of the debate, and distils the core argumentative threads under distinct scenarios for AI progress. Overall, it builds a detailed and robust case for the plausibility of short TAI timelines.

Context and motivations. This report is intended for a broad audience, including researchers and decision-makers in AI safety and governance, as well as policy-makers and forecasters with an interest in transformative technologies. Perspectives vary widely within these communities, from those who dismiss short TAI timelines as implausible or even absurd to those who are convinced of their inevitability.

I hope that this report proves valuable regardless of where on this spectrum of belief the reader falls. In particular:

- *For the sceptical reader*, the arguments laid out might challenge some of her assumptions, and perhaps even convince her that the prospect of short timelines should be taken seriously by decision-makers. At the very least, it should help her better understand and engage with her opponent's case.
- *For the reader who already takes short TAI timelines very seriously,* this report offers a consolidated body of evidence she can draw upon and reference in support of her arguments.
- *For all readers*, it highlights key uncertainties in the debate, and identifies areas whether further research could be useful, helping them refine their own views and the directions of their future work.

In this sense, I see the primary value of this report as lying less in the overarching conclusion – that *short TAI timelines are plausible*, a stance that is not unusual in this space – and more in its utility as a resource for understanding, engaging with, and building upon the existing timelines debate.

Having said this, I do think there is a strong case for the plausibility of short timelines to be found here. The evidence outlined in this report was not selected with the intention of reinforcing any one particular narrative, and the conclusion is an honest reaction to the evidence that was available. Ultimately, when we aggregate and examine the best arguments on both sides of the debate, the development of TAI appears to be a realistic outcome of the next ten years of capabilities progress, at minimum passing the bar for consideration by decision-makers in safety and governance contexts.

Adaptability of the arguments. The state of the timelines debate is constantly changing with each new development in the AI field. Therefore, rather than simply laying out the existing evidence for and against short TAI timelines, this report also seeks to construct arguments that are robust to the shifting tide of discourse and the emergence of new evidence.

Crucially, the analysis of the report illustrates that even if one route fails to get us to transformative AI by 2035, there are several other ways we could plausibly end up in a short timeline. (So, we should avoid overindexing our timeline beliefs on any *one* possible obstacle for development of transformative AI.)

For example, even as I write this, news has just broken out that the performance gains from scaling training compute in AI systems may be beginning to break down. If this is accurate, my report equips the 'believer' in short timelines with a response; it presents a number of different ways we might still develop TAI by 2035 even if the contributions from compute scaling diminish during this period.

> DISCLAIMER
>
> **I stopped making substantive edits to the content of this report in December 2024.** This means that any developments in the AI field that have occurred after December 2024 have not been captured here. Moreover, some news that broke when the report was undergoing its final revisions at the end of 2024 (such as the whisperings of a breakdown in scaling trends, as linked above, and the launch of OpenAI's o3 model) has only been lightly integrated into the report: I've typically highlighted these news items where they are relevant to the local arguments, but their consequences for the overarching narrative of the report (including our level of confidence in its conclusion) have not always been spelled out in detail.

How to read this report. I would encourage all readers to begin with the 1-page summary, followed by the 13-page detailed overview. The latter should provide sufficient information to then guide the reader towards the sections of the report that will be most of interest to her. However, the individual sections of this report are not standalone items, and it will often be necessary to read earlier content to understand the terminology and arguments referred to later on.

Broadly, I expect that:

- *If you don't have much pre-existing knowledge on AI safety, governance, or capabilities*, then you are likely to benefit from a close reading of the Introduction and Chapters 1 and 2. Here, the necessary background for understanding the timelines debate is laid out, a focus on two key

'mechanisms' of capabilities progress is motivated, and the arguments for and against those mechanisms resulting in a short TAI timeline are laid out and evaluated in detail.

- *If you already have a background in AI safety, governance, or capabilities research*, Chapter 3 might be of more interest to you than the earlier chapters (though the earlier chapters should at least be skimmed). Here, I construct and briefly describe seven meaningfully distinct, plausible scenarios with short TAI timelines. This is where the report's main contributions to the scenario planning literature lie, and may provide useful inputs for other research in this direction.

The Conclusion is also likely to be of interest to the majority of readers, serving as both a detailed explanation of how we should react to the evidence I have laid out, as well as a defense of the claim that short TAI timelines are plausible. At the end of this chapter, I also highlight key uncertainties and present a 'bounty list' of areas for future research; these elements may be especially useful to any readers who wish to build upon the findings of this report or integrate them into their own work.

Convergence Analysis

Zershaaneh Qureshi
Researcher at Convergence Analysis

Email address:
zershaaneh.qureshi@convergenceanalysis.org

LinkedIn profile:
https://www.linkedin.com/in/zershaaneh-qureshi-8744131b4/

Acknowledgements

Thank you to Justin Bullock, Elliot McKernon, Daan Juijn, Jakub Growiec, Tom Davidson, Armand Bosquillon de Jenlis, Anson Ho, and Seth Blumberg for feedback on this report.

1-page summary of report

This report explores pathways through which AI could develop transformative capabilities within the next ten years (i.e. by 2035), referred to here as 'short timelines' to transformative AI (TAI). It focuses on two primary mechanisms for rapid AI capabilities progress – *compute scaling* and *recursive improvement* – which play central roles in the most influential stories of short TAI timelines. The detailed analysis of this report culminates in a case for the plausibility of short timelines.

Compute scaling. The history of AI development indicates that AI capabilities can be improved by increasing (effective) compute – and we've observed fast growth in this input, fuelled by increases in microchip density, hardware efficiency, algorithmic progress, and investment. Some experts believe these trends will persist over the next decade. If so, they could result in TAI before 2035.

Sceptics argue that this pathway will soon face challenging bottlenecks (concerning e.g. data, investment, power, and limitations of traditional LLMs) that would slow progress. However, even if compute scaling becomes seriously bottlenecked on something before TAI arrives, other mechanisms – such as recursive improvement – could still achieve enough traction to produce TAI within the next ten years.

Recursive improvement. If AI systems are deployed to automate AI R&D, they could initiate powerful feedback loops in the AI field. Some argue that this would not only break bottlenecks to compute scaling, but drive exponential or even super-exponential growth in AI capabilities, resulting in the arrival of TAI (and perhaps even more advanced systems) before 2035.

Sceptics argue that the effects of these feedback loops would be counteracted by increasing bottlenecks and diminishing returns on effort as low-hanging fruit in capabilities improvements is exhausted. They also highlight constraints which would limit the size or speed of each capabilities improvement 'step'. However, even if recursive improvement cannot drive exponential growth in AI capabilities, it could still enable fast enough progress to achieve TAI by 2035.

Short timeline scenarios. On examination, it seems that there are many different routes through which TAI could arrive by 2035. To illustrate this, I generate and describe seven plausible scenarios with short TAI timelines. In five of these, represented in Figure 3.1, progress is based on compute scaling and/or recursive improvement; the other two highlight pathways to short TAI timelines which *don't* significantly rely on these mechanisms. The existence of a plurality of plausible short timeline scenarios strengthens the evidence base for short timelines.

Detailed overview (13 pages)

Purpose and audience

This report is intended for a wide audience, including researchers and decision-makers in AI safety and governance, as well as policy-makers and forecasters with an interest in transformative technologies.

The motivations for this report, as well as the best ways to use it, are detailed in the earlier section entitled 'Purpose of this report'. In broad terms: it is a resource for better understanding, engaging with, and building upon, the debate about timelines to 'transformative AI'. Overall, it should encourage readers to take seriously the prospect of such systems being developed within the next ten years. It also provides a consolidated body of evidence for the plausibility of this outcome, highlighting key uncertainties and areas for further investigation.

Background

I define 'transformative AI' (TAI) as AI systems which are capable of transforming society to an extent comparable to the industrial or agricultural revolutions. AI capabilities levels that might be considered 'transformative' in this sense include artificial general intelligence, human-level machine intelligence, superintelligence, and other familiar notions from the literature on advanced AI. (For a more detailed enumeration of systems that could qualify as TAI, see the subsection of this report titled 'What capabilities could constitute TAI?'.)

The date of the arrival of the first TAI systems[1] is of great strategic relevance in the context of AI safety and governance. It determines the urgency of action and the specific policies, safeguards, and risk mitigation measures that can feasibly be implemented before society is radically transformed.

With this in mind, this report continues Convergence Analysis' exploration of the *timeline to TAI* as a strategic parameter for AI scenarios. It follows my previous article on Timelines to Transformative AI, which mapped out the current landscape of TAI timeline predictions and examined the trends emerging from that landscape.

In the present report, I now explore pathways through which AI could develop transformative capabilities within the next ten years. I describe scenarios in which TAI is developed in the next ten years as exhibiting 'short TAI timelines'.

I focus especially on two key mechanisms for AI progress – compute scaling and recursive improvement – which play central roles in some of the most influential stories of fast capabilities development. For each mechanism, I

[1] Here, I'm specifically interested in the date of their initial arrival in a lab setting. There are further questions of strategic importance around the timeline for deployment and diffusion, but I do not address these in this report.

dedicate a chapter to examining arguments for and against its producing a short TAI timeline.

I also devote some time at the end of the report, in Chapter 3, to describing seven distinct scenarios in which TAI arrives in the next ten years. Under five of these scenarios, progress over the next decade is driven by some combination of compute scaling and/or recursive improvement.[2] The other two scenarios highlight pathways to short TAI timelines which *don't* significantly rely on these popularly discussed mechanisms.

Two key mechanisms for fast AI capabilities progress

The most popular stories of short TAI timelines typically appeal to at least one of the following mechanisms for progress as the primary basis for fast capabilities improvements in AI R&D over the next decade:

- **Compute scaling.** In many short timeline stories, compute plays a central role in driving AI capabilities progress over the next decade. In particular, AI systems within the current paradigm are argued to become increasingly capable as the amount of compute used to train them is increased.[3] This implies a short timeline if AI systems can be fed with enough compute to reach TAI by 2035.
- **Recursive improvement.** This is a broad category of mechanisms for AI capabilities improvement. It includes any positive feedback loops through which there are repeated improvements to the ability to improve AI systems. This includes, for example, the investment feedback loops that are currently supporting AI development. However, in the context of short timelines, the *most popular* stories of recursive improvement posit the future emergence of what I call 'direct' feedback loops: AI systems *themselves* begin to drive capabilities improvements by contributing to AI R&D, and get better at doing so with each subsequent improvement. This implies a short timeline if AI systems can be deployed to make sufficiently fast improvements to AI capabilities within the next decade.

This report primarily seeks to better understand these two mechanisms for fast AI capabilities progress (rather than, say, comprehensively charting out all arguments in the literature for and against short TAI timelines). However, it does briefly touch on other routes of capabilities progress.

Compute scaling

In Chapter 1 of this report, I explore compute scaling as a mechanism for fast AI capabilities progress. I begin by outlining the role of compute in recent capabilities progress, and how continued compute scaling might produce TAI within the next ten years. I then go on to consider why it might fail to do so.

[2] These five scenarios are represented by a scenario tree, Figure 3.1, which is also included later in this overview section.

[3] Very recent evidence (e.g. from OpenAI's o1 model) suggests that AI systems also become increasingly capable with increased run-time compute. In light of this, 'run-time compute scaling' is increasingly featuring in stories of future AI capabilities progress, and is discussed in Chapter 1 of this report.

Historically, however, training compute has been the focus in most arguments for short timelines via compute scaling. There is also less empirical data on the scaling relationship between run-time compute and capabilities, and the implications of results like those very recently published by OpenAI are not yet fully understood. Because of this, this report focuses primarily on the prospect of scaling training compute – but the existence of a second route of compute scaling is taken to strengthen the overall case for expecting short TAI timelines, and provide a potential line of response to some of the sceptic's objections.

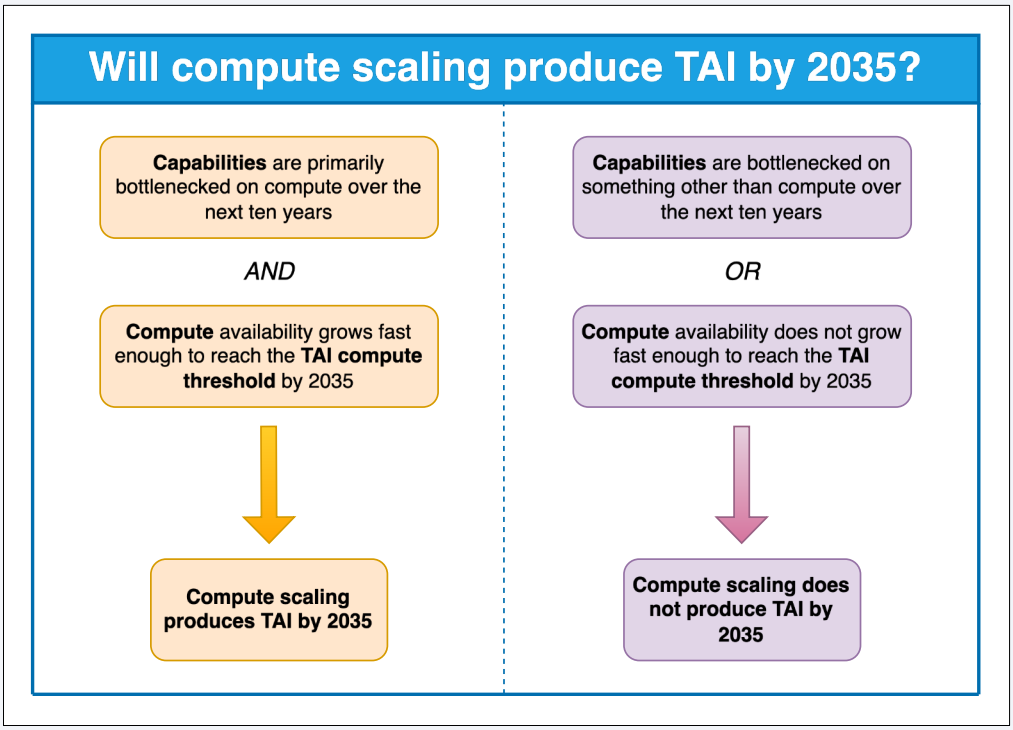
Through these arguments, it becomes apparent that compute scaling pathways of AI capabilities progress will *eventually* come up against difficult bottlenecks. However, meaningful debate remains over the extent to which these challenges will emerge over the next ten years, how difficult they will be to address or sidestep, and – if they do become prohibitive – whether TAI will have already been achieved by the point at which this happens.

Short timelines via compute scaling

Arguments that compute scaling will result in TAI within the next ten years are often based on the two following claims:

i. Compute will remain the primary driver of capabilities progress in LLMs on the pathway to TAI; that is, progress will not be significantly bottlenecked on factors other than compute. (I describe this view as a 'compute-centric variant of the *scaling hypothesis*'.)

ii. The compute that can be used to train frontier LLMs will grow fast enough for a TAI-scale training run to be achieved by 2035.



As I lay out in the subsection entitled 'What is compute scaling, and how could it produce a short timeline?', extrapolating trends from past AI progress provides some support for both of these claims.

- Firstly, progress in neural network capabilities has so far been in line with the compute-centric scaling hypothesis: increasing the compute used to train a system has resulted in improved performance. Some take these historical observations to suggest that neural network capabilities will continue to increase in a predictable way as training compute increases, following *empirical scaling laws* (e.g. those from OpenAI and DeepMind). This interpretation of the data can be provided in support of (i).

- Secondly, over the past few decades, we've seen fast and consistent growth in the compute used to train frontier AI models (see e.g. Epoch's data on historic compute trends in machine learning). We've simultaneously observed consistent improvements in many of the *underlying inputs* to both physical and effective compute growth (such as microchip density, as captured by Moore's Law, as well as investment, hardware price-performance, and algorithmic efficiency). Extrapolating from these historic trendlines provides some support for the plausibility of (ii). (Of course, the plausibility of (ii) also depends on how much compute will actually be required to train TAI; this thread of the argument is omitted in this summary, but discussed in some detail in the full report.)

If we take this historic data into consideration when predicting the arrival of TAI, short timelines appear to be a real possibility.

Counterarguments to short timelines via compute scaling

It's not clear how far we can extrapolate from these past trends to draw conclusions about the future of AI development. Even if compute has been the primary driver of capabilities progress in LLMs *so far*, it may not continue to play this role in future; at some point, progress may become bottlenecked on some other factor. Similarly, some of the sustaining forces behind historic trends of compute growth are likely to *eventually* break down. If any of these trends break down or lose momentum within the next ten years, the likelihood of a short TAI timeline would be reduced.

The sceptic can therefore object to both (i) and (ii). In 'Why compute scaling might not produce a short TAI timeline', I consider objections on both points separately, under two broad categories of sceptical argument.

Category 1: 'Capabilities progress will be bottlenecked on something other than compute'. This type of objection involves a rejection of assumption (i).

Sceptics in this category argue that, at some point in the next ten years of AI development, capabilities progress will become bottlenecked on inputs *other* than compute, which might be considerably harder for developers to deal with.

This potentially includes:

- *Data (quality/quantity)*. We might run out of high quality data to train large-scale AI systems on before TAI is developed. If solutions cannot be found in the form of self-play/synthetic data generation, this could be a significant barrier to developing systems with transformative capabilities.
- *Algorithms*. Traditional LLMs may struggle to achieve the levels of generality required for TAI, given their historically weak performance on benchmarks such as ARC; if so, substantial algorithmic progress or even an AI paradigm shift may be needed before TAI can be developed.[4]
- *Training environment (i.e. the design of the overall ecosystem of training*

[4] Note that this line of argument against short timelines has been somewhat undermined by the recent breakthrough performance of OpenAI's o-series models on the ARC benchmark. Details of what to make of these developments in the context of the debate over generality in current AI systems can be found in the relevant section of Chapter 1.

data and training algorithms – not just quality/quantity of data). It might be very difficult to design a training environment which actually incentivises the development of transformative capabilities; if so, the need for trialling many training environments may delay the arrival of TAI. (See Richard Ngo on the 'hard paths hypothesis'.)

- *Embodiment*. The emergence of certain capabilities relevant to TAI might rely on an AI system being physically or virtually embodied; if so, significant further R&D may be required in this direction.

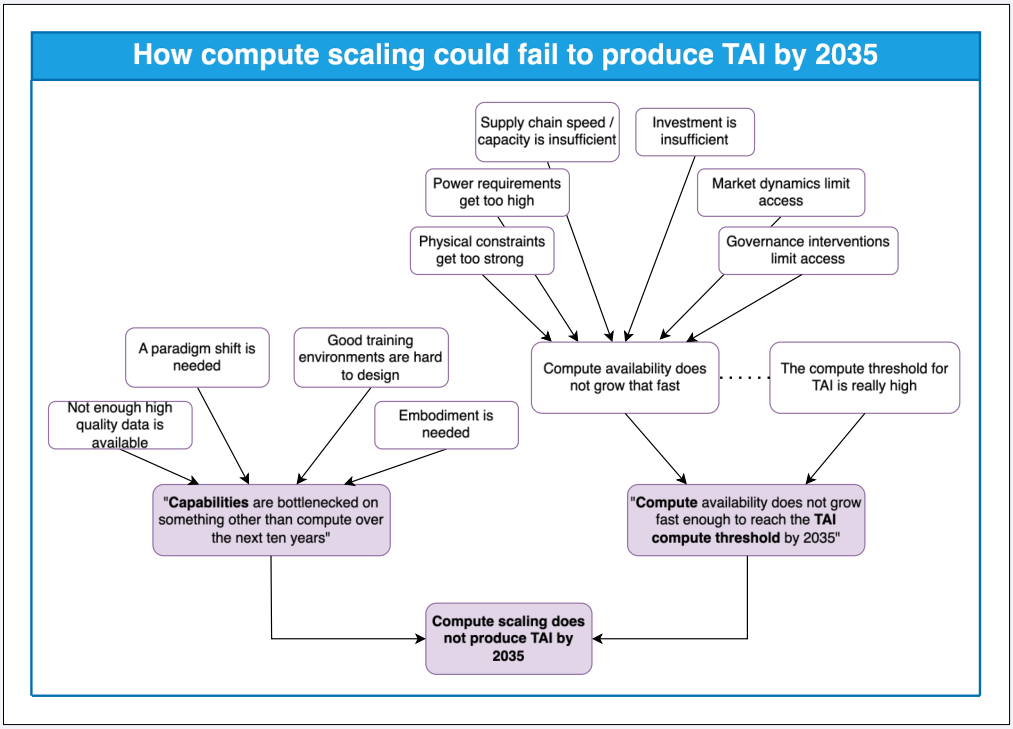
If any of these factors overtake compute as the primary bottleneck for AI capabilities progress, dealing with them could add significant delays to the arrival of TAI.

Category 2: 'Compute growth will not be fast enough for TAI by 2035'. This type of objection involves a rejection of assumption (ii).

Sceptics in this category typically argue that growth in the compute used to train frontier AI systems will be considerably slowed down over the next decade, by, for example:

- *Physical limitations* on the density of chips
- *Power requirements* of TAI-scale training runs being unachievable
- *Investment requirements* of TAI-scale training runs being unachievable
- Limitations on *chip manufacturing capacity* and the speed of the *supply chain*
- *Market dynamics* preventing a single actor from obtaining a high enough proportion of total available compute to train TAI
- *Governance interventions* restricting compute access

Sceptics in this category might also point out that the level of compute required for TAI-level capabilities could be so high that it's simply out of reach within the next decade, without much faster compute growth than is realistic. In support of this claim, we could gesture to the *uncertainty* over how much compute would be needed for TAI-scale training run, the *long tails* of the distribution of TAI compute requirements, and the *difficulty of supplying a meaningful upper bound* for compute requirements here.



Reflections. Although compute is, and will likely continue in the near term to be, an important driver of AI capabilities progress, it's not clear exactly how far compute scaling can take us over the next decade. It seems likely that, at some point, other bottlenecks will emerge and have a slowing or limiting effect on AI capabilities progress. If this happens *before* TAI has arrived, it might seriously reduce the likelihood of a short TAI timeline.

Recursive improvement

Even if the identified challenges for compute scaling *are* poised to slow down capabilities progress on the pathway to TAI, this doesn't mean that short timelines are off the table. There are other mechanisms through which AI capabilities progress could still be fast enough for TAI to arrive by 2035. In Chapter 2, I examine *recursive improvement* as a broad category of such mechanisms.

I begin this chapter by outlining different types of recursive improvement, and how they could result in TAI arriving within the next ten years. (In 'Overcoming scaling bottlenecks via recursive improvement', I note especially the potential for certain recursive improvement dynamics to break some of the previously identified bottlenecks to compute scaling, reinforcing compute scaling pathways to TAI.) I then go on to examine arguments that these mechanisms might fail to yield a short timeline.

Reflecting on this discussion, I note that if a period of (what I call 'direct') recursive improvement does begin in the next few years, it's hard to argue that this would not result in the arrival of TAI by 2035. It's reasonable to argue that recursive improvement wouldn't necessarily lead to a *sustained period of acceleration* in capabilities (e.g. exponential or super-exponential trajectories

of improvements) – but even so, the believer in short timelines has room to argue that capabilities improvements would still be fast enough to produce a short TAI timeline.

Short timelines via recursive improvement

I broadly define 'recursive improvement' as any iterative process characterised by feedback loops through which there are *repeated improvements to the ability to improve AI*. (See my introduction to 'What is recursive improvement, and how could it produce a short TAI timeline?' for further details here.)

Direct recursive improvement

I focus my attention on positive feedback loops which are mediated *directly* by AI systems. I call these 'direct' feedback loops for AI capabilities progress.

The section titled 'Direct recursive improvement' outlines the 'AI R&D type' direct feedback loops introduced by AI systems which can automate significant parts of AI R&D. Here, I frame things around the idea of 'automated workers' for AI R&D: AI systems which can perform all or most of the tasks typically performed by a human researcher or engineer (with minimal human supervision/prompting) and can thereby effectively act as drop-in replacements for those humans. (However, I also note that there are alternative ways in which AIs could contribute to AI R&D.)

Thus conceived, the main thrust of the argument for direct recursive improvement is as follows: once automated AI R&D workers are developed, large numbers of these systems are deployed in parallel, vastly increasing the total number of human-equivalent hours being spent on making AI capabilities improvements. This first generation of automated workers drives the development of a second generation which is even more capable than the first at AI R&D, and therefore *even better equipped to make improvements* to subsequent generations of models than its predecessors were. A cycle of positive feedback emerges in which AI capabilities improve, step by step.

With this story in mind, I go on to outline the possible trajectories of step-by-step capabilities improvements that could result from AI R&D automation. I also briefly highlight some support from prominent empirical research and quantitative models for the claim that direct recursive improvement would underpin an accelerating trajectory of capabilities improvements.

Indirect recursive improvement

Direct recursive improvement dynamics cannot be sustained without increased inputs from what I call the 'indirect' feedback loops operating in the background. In the subsection of Chapter 2 titled 'Indirect recursive improvement', I highlight a number of much broader societal feedback loops which play a crucial role in AI capabilities progress. These include economic feedback loops (driven by reinvestment of capital into AI R&D), scientific feedback loops (driven by advancements in scientific tools and methods) and

political feedback loops (driven e.g. by competitive pressures/race dynamics). Without these background processes providing sufficient resources and motivations for improving AI capabilities, any period of capabilities growth via direct recursive improvement would likely plateau.

Although indirect feedback loops could be powerful mechanisms for progress in their own right, I focus in this chapter on arguments for short TAI timelines which specifically invoke direct recursive improvement.

Short timelines

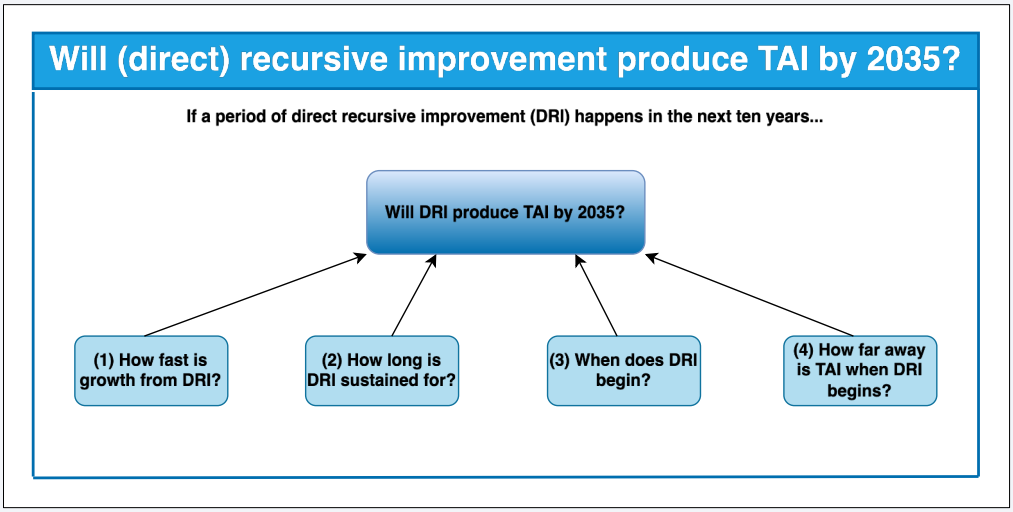
For the sake of this chapter, I assume that the AI field will eventually reach a capabilities threshold at which direct recursive improvement can begin. With this granted, there are four further questions which then determine whether the ensuing period of direct recursive improvement (DRI) will result in a short timeline to TAI:

(1) How fast is the capabilities growth resulting from the feedback loops at play? (i.e. what is the shape of the trajectory of recursive improvements?)

(2) How long is this period of DRI sustained for?

(3) When does this period of DRI begin?

(4) How far away is TAI when this period of DRI begins?



The believer in short timelines via recursive improvement will argue that the answers to these four questions are favourable towards TAI arriving within the next ten years. That is: direct recursive improvement dynamics will be fast enough, sustained for long enough, and kick in soon enough to cross the distance to TAI by 2035.

There is especially interesting discussion in the literature over the first two points. Some have suggested that direct recursive improvement dynamics could enable exponential or super-exponential modes of capabilities growth that would continue until a 'singularity' in AI development is reached.

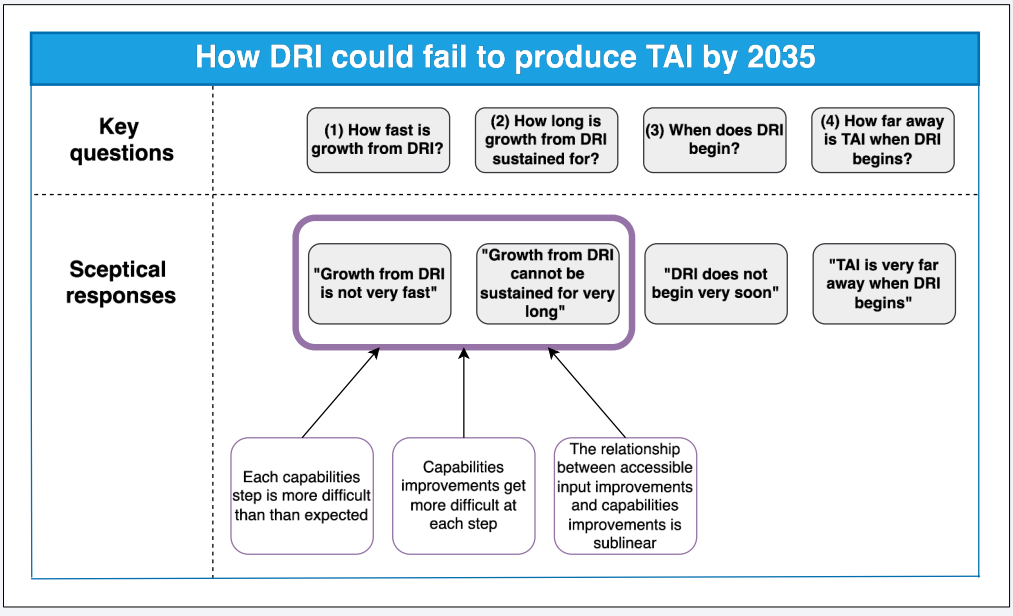
Counterarguments to short timelines via recursive improvement

The sceptic of short timelines via (direct) recursive improvement can level objections in response to any of the four questions listed above.

In 'Why recursive improvement might not produce a short TAI timeline', I focus on sceptical arguments which respond to questions (1) and (2). That is: I'm interested in reasons to think that a recursive process of capabilities improvement 'steps' would either not be very fast, or could not be sustained for a long time. I examine three such arguments in detail:

- **'Each capabilities improvement step is more difficult than expected'.** Each capabilities step will be harder to make than has been imagined by those who expect short timelines via recursive improvement, due to difficult situational requirements that they have not adequately factored in.
- **'Capabilities improvements get more difficult at each step'.** Each capabilities step will be harder to make than the previous one, due to the increasing effect of bottlenecks (e.g. demands on compute and energy), as well as diminishing returns on effort.
- **'The relationship between accessible input improvements and capabilities improvements is sublinear'.** Even if recursive improvement dynamics enable certain *inputs* to AI capabilities to improve in big/fast steps, AI capabilities themselves might only improve in small/slow steps. (This may be best understood as a variant of the previous bullet point.)

If any of these arguments are taken seriously, the consequence (according to the sceptic) is that capabilities improvement 'steps' will be small/slow, or will soon reach a plateau beyond which further improvements cannot feasibly be made. This would call into question whether TAI could actually be achieved by 2035 under a direct recursive improvement scenario.



Reflections. It seems there are many factors which could plausibly constrain the trajectory of AI capabilities progress during a period of direct recursive improvement. Moreover, although no counterarguments from this chapter are decisive, they at least provide reasons to *doubt* claims that direct recursive improvement would enable a sustained period of exponential or super-exponential AI capabilities growth.

However, as I argue in 'Who wins the tug of war?', it's not easy for the sceptic to win this argument against the believer in short timelines:

- Those who expect short timelines via direct recursive improvement can, and sometimes do, incorporate many of the constraining factors detailed above into their models. They just don't believe that this will slow progress down *by enough* to prevent a short TAI timeline, or mean that capabilities improvements will plateau *any time soon*.
- Exponential or super-exponential trajectories of AI capabilities improvements are probably not necessary for a short TAI timeline to be realised. If the emergence of direct recursive improvement dynamics simply yields a one-time step change in the rate of capabilities improvements, or empowers the field to overcome bottlenecks to compute scaling and thereby helps to sustain current rates of progress, this might still be sufficient for producing TAI by 2035.

There are, of course, other lines of argument the sceptic can pursue instead. For example, she can target questions (3) and (4) on the list above, arguing that direct recursive improvement dynamics will not kick in any time soon (perhaps not even within the next ten years) or that TAI-level capabilities are just extremely far away. I discuss these options in the subsection titled 'Other objections'.

Seven scenarios with short TAI timelines

In Chapter 3 of this report, I synthesise the core elements and argumentative threads of previous chapters in a more concrete way. I do this by outlining a series of scenarios with short TAI timelines which each seem (at least somewhat) plausible in light of earlier reflections, but which differ on certain core assumptions.

First, I characterise five plausible short TAI timeline scenarios in which capabilities progress is driven by some combination of compute scaling and/or recursive improvement. ('Five scenarios based on compute scaling/recursive improvement'.)

As explained in the subsection entitled 'Scenario generation methodology', the five scenarios in this set are determined by the differing values they assign to the following parameters:

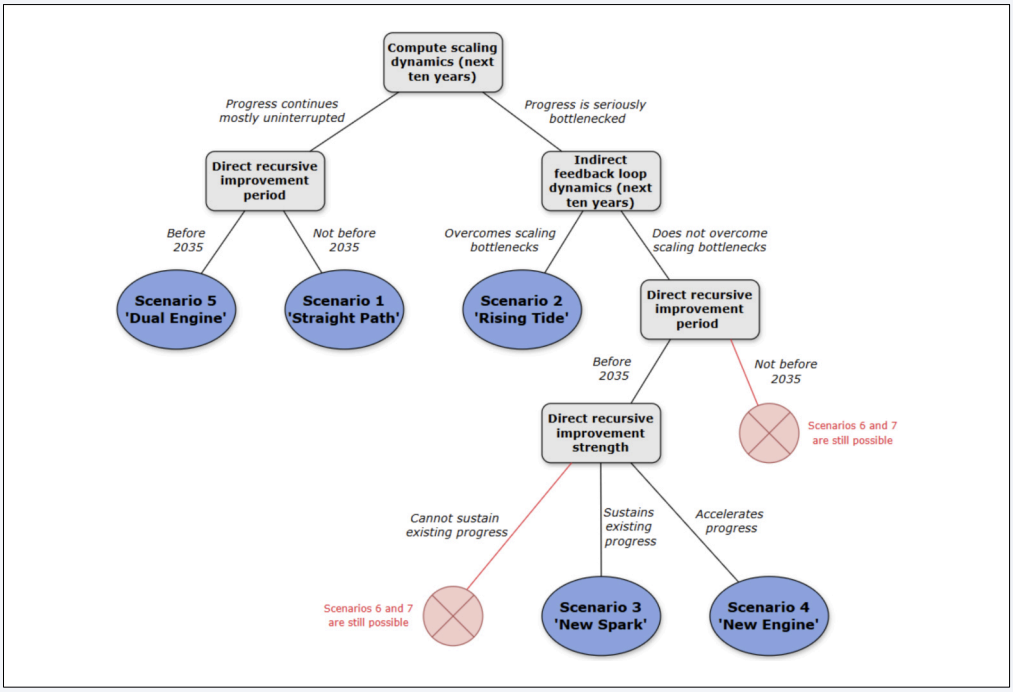
- **Compute scaling dynamics (next ten years).** Will compute scaling continue as before over the next ten years, or will this route of capabilities

progress soon get seriously bottlenecked on something (be it data, paradigmatic limitations, power requirements, the supply chain, or anything else)?

- **Indirect feedback loop dynamics (next ten years).** If compute scaling encounters serious bottlenecks, will indirect feedback loops gain enough traction in the next ten years to overcome them?

- **Direct recursive improvement period.** Will a period of direct recursive improvement begin within the next ten years?
  - **Direct recursive improvement strength.** If a period of direct recursive improvement does begin within the next ten years, how strong will this be? Will it be strong enough to accelerate existing rates of progress, merely strong enough to sustain existing rates of progress in the face of growing bottlenecks to scaling, or not even strong enough to overcome those bottlenecks?

These parameters form the nodes of the tree below, which characterises the decision process through which these five short timeline scenarios are generated.



SCENARIO 1

'Straight Path'. *Compute scaling just works.*

Compute scaling with the current paradigm continues to yield results and does not become *seriously* bottlenecked in the next ten years.[5] There are problems to solve along the way (e.g. on the side of data or algorithms), but there are quick fixes available (e.g. synthetic data generation[6] works well, and unhobbling leads to easy improvements in LLM generality). Direct recursive improvement does not kick in at any point, but doesn't need to; compute scaling is enough to produce TAI by 2035.

[5] Or, if it does get seriously bottlenecked, another form of compute scaling (e.g. with run-time compute rather than training compute) works just fine. I don't mention this option explicitly in my scenarios, but take it to basically be a variant of what I call 'compute scaling' here. Of course, it only applies in cases where the bottleneck to compute scaling is not a *lack of physical compute*.

[6] Note: I do not consider synthetic data generation alone as sufficient for underpinning what I call a period of 'direct recursive improvement' in Chapter 2. I do, however, accept that AIs which generate data could bring about a much more restricted (and therefore weaker) form of the same dynamic. This will become clear in Chapter 2.

SCENARIO 2

'Rising Tide'. *IRI breaks bottlenecks.*

Compute scaling gets seriously bottlenecked on something in the next ten years (e.g. at some point, developers just can't afford enough compute to continue scaling systems up). However, indirect feedback loops in the background gain traction over the next ten years. (For example, AI systems attract some capital which can be reinvested into procuring more compute, the scaled-up AI systems perform better and attract even more capital, and so on.) This helps to lift capabilities progress out of a plateau. Direct recursive improvement *could* also kick in at some point, but doesn't need to; compute scaling plus indirect recursive improvement is enough to produce TAI by 2035.

SCENARIO 3

'New Spark'. *Moderate DRI sustains progress.*

Compute scaling gets seriously bottlenecked on something in the next ten years. Indirect feedback loops do not gain sufficient traction to lift capabilities progress out of this plateau. However, a period of direct recursive improvement soon kicks in. It's strong enough to sustain current rates of capabilities progress. Systems are near enough to TAI-level capabilities at the time that direct recursive improvement kicks in for TAI to be produced by 2035.

SCENARIO 4

'New Engine'. *Strong DRI accelerates progress.*

Compute scaling gets seriously bottlenecked on something in the next ten years. Indirect feedback loops do not gain sufficient traction to lift capabilities progress out of this plateau. However, a period of direct recursive improvement soon kicks in. It's strong enough to accelerate capabilities progress. (For example, there could be a one-time step change in the rate of capabilities progress, or a sustained period of continuous acceleration.) Even if systems are far away from TAI-level capabilities at the time that direct recursive improvement kicks in, this doesn't matter; direct recursive improvement leads to such fast (and/or prolonged) capabilities progress that TAI is still produced by 2035.

SCENARIO 5

'Dual Engine. *Joint compute scaling + DRI accelerates progress.*

As in Scenario 1, compute scaling with the current paradigm continues to yield results and does not become *seriously* bottlenecked on anything in the next ten years. There are problems to solve along the way, but there are quick fixes available. Direct recursive improvement also kicks in within the next ten years. Even if systems are far away from TAI-level capabilities at the time that direct recursive improvement kicks in, this doesn't matter; direct recursive improvement plus continued compute scaling leads to such fast (and/or prolonged) capabilities progress that TAI is still produced by 2035.

In 'Have we missed anything important?', I then outline two other scenarios which do not significantly rely on either compute scaling or direct recursive improvement as primary mechanisms for AI capabilities progress over the next decade, but could still yield a short TAI timeline. These both point to a new approach to AI development which, once adopted, enables TAI to be produced relatively quickly.

- **Scenario 6: 'LLM Hybrid'.** A hybrid architecture is developed which combines LLMs with a form of symbolic reasoning or new learning methods. This displays much higher levels of generality than the current paradigm. Relatively minor or fast improvements to this hybrid paradigm are sufficient to achieve a form of TAI by 2035.
- **Scenario 7: 'Intelligent Network'.** Before 2035, many systems, each with narrow capabilities, are composed together in a network (e.g. in the style of Drexler's Comprehensive AI Services). The combination of these systems' individual capabilities constitutes a genuinely transformative composite system.

Reflections. Some of these scenarios might seem less plausible than others. I do not favour any one scenario as being especially likely to occur. However, at the end of this chapter, I argue that the very existence of this plurality of routes through which TAI could feasibly be achieved by 2035 is noteworthy, and should strengthen our overall degree of belief in short timelines. This argument is picked up again in the Conclusion.

I also note that the specific pathway we end up taking to TAI (and not just the timeline) is of strategic importance. In 'Which scenario?', I speculate about how the scenario we are in influences the type of transformative system that arrives first, how far we will have surpassed TAI by (if at all) in 2035, and the warning signs (if any) we can expect to have along the way.

Conclusion

In the 'Conclusion' section of this report, I argue for the plausibility of short TAI timelines on the following grounds:

- There are several different routes through which TAI could conceivably be achieved in the next ten years. If AI capabilities progress is slow or begins to plateau on one route, other mechanisms could soon kick in through which TAI might still quickly be achieved.
- Short timeline scenarios are compatible with a variety of different background assumptions (for example, about scaling, the current paradigm, and the strength of different drivers and restraints of AI capabilities progress).
- The body of evidence in support of short timeline scenarios is rapidly growing. With new developments in the AI field, the state of this debate appears to be shifting more and more towards short timelines as a likely outcome of capabilities R&D efforts.

In ‘What now?’, I go on to note key areas of uncertainty over the arguments I have laid out, as well as areas of strategic importance which warrant further exploration. This, alongside the other takeaways of this report, motivates some important questions for further research, which are captured under a ‘bounty list’ at the end of this document.

Introduction

This report is part of a broader research project seeking to answer the following questions:

How could AIs with 'transformative' capabilities be developed by 2035? And what might these 'transformative' systems actually look like?

In this report, I examine the plausibility of AI capabilities reaching transformative levels within the next ten years. I do this by exploring the different routes through which progress could be especially fast, focusing primarily here on two key mechanisms for AI capabilities progress: *compute scaling* and *recursive improvement*. After individually examining the debates over these mechanisms, I go on to distil the core arguments under seven distinct scenarios of fast AI capabilities progress.

Note: this research is *not* about progress *beyond* the first transformative systems (e.g. takeoff from AGI to superintelligence). For example, although recursive improvement is most commonly invoked in stories of rapid AI progress beyond the arrival of AGI, I am primarily concerned with the potential for recursive improvement to take us to AGI or similar. Definitions of AGI and other relevant conceptions of advanced AI are provided in 'What capabilities could constitute TAI?'.

This introductory section is intended to lay the groundwork for the report. I begin by providing relevant background and clarifying some definitions. I'll then introduce a simple model of AI progress that will be useful to refer back to in subsequent chapters of the report, and identify several ways in which AI progress could be especially fast. With this groundwork in mind, I'll carve out the territory for the rest of this report.

Background on *AI Clarity*'s previous work

A previous article from *AI Clarity* introduced scenario planning for AI x-risk, including a discussion of strategic parameters as a tool for analysing AI scenarios.

One key strategic parameter for AI scenarios is the ***timeline to transformative AI*: the date at which we develop AI systems which are capable of transforming society to an extent comparable to the industrial or agricultural revolutions.**

As part of *AI Clarity*'s scenario planning efforts, I previously published a detailed investigation into the current landscape of predictions of the timeline to transformative AI (TAI), in which I found that:

- The majority of recent median predictions for the arrival of TAI (elicited

from experts, forecasters, and the public) fall within the next 10 to 40 years.

- Over the last few years, and across most groups, people generally seem to be updating their beliefs in the direction of shorter timelines to TAI.

This report will continue my examination of TAI timelines as a strategic parameter. I now aim to tell a more complete story, elucidating the *why?* and *how?* of different perspectives on TAI timelines. Specifically, given a 'short timeline' to TAI, I want to understand the conditions of the world that could have led us there, and the arguments for and against such conditions being realised.

What do we mean by 'short timelines'?

By highlighting 'short timelines', I mean to focus my attention on the lower end of the 10-40 year spectrum I identified in my previous article. From my observations there, timeline predictions of ten years or less from today (i.e. with TAI arriving before 2035) fall around the earliest dates of the current range of popular opinion amongst experts and forecasters, but are being taken increasingly seriously as of the last few years.[7] It's this realm of TAI timeline opinions which are *not quite* mainstream but gaining real traction – and require very fast action if accurate – that I want to investigate.

Taking my previous observations as a rough guide, let's define a 'short TAI timeline' as one in which *TAI is developed within 10 years of today*.

A note on this definition. The milestone of primary interest to me here is when TAI-level capabilities arrive – not when systems with such capabilities are *deployed* and actually transform society, or when the *impact* of AI (for example, on the economy) reaches any particular threshold. However, the deployment and impact of pre-TAI systems will sometimes be relevant to the discussion, inasmuch as such factors could influence the pace of the development of TAI-level capabilities.

This focus on 'TAI-level capabilities' motivates another (rough) definition.

What capabilities could constitute 'TAI'?

The term 'TAI' may be used to refer to any AI system capable of producing certain *impacts*; absent further explanation, the term is silent on the exact nature of, or *capabilities* exhibited by, such systems.

For my present purposes, I won't rigidly define 'TAI-level capabilities', but it will be instructive to highlight a few examples of capability levels that *could* be considered 'transformative':

- **Artificial general intelligence (AGI).** (This is the most common characterisation of a transformative level of capabilities, but is variably operationalised.) In particular, Levels 3, 4, or 5 of Google DeepMind's

[7] Key sources from the previous article include: expert surveys conducted by AI Impacts (compare the results of the 2023 survey to those of the 2022 survey to see the recent shifts in opinion); Metaculus' community predictions for the arrival of AGI; results from the Forecasting Research Institute's Existential Risk Persuasion Tournament; quantitative forecast models from Ajeya Cotra and Epoch; predictions made by AI lab leaders such as Sam Altman and Dario Amodei; and AI experts such as Yoshua Bengio and Geoff Hinton who have recently revised their timelines downwards.

All sources discussed in that article were compiled into a table which can be accessed here.

tiered AGI classification system could correspond to TAI. These levels capture AI systems which, "across a wide range of non-physical tasks"[8], consistently:

- (Level 3) Fall in the 90th percentile of human performance.
- (Level 4) Fall in the 99th percentile of human performance.
- (Level 5) Outperform 100% of humans.

Alternatively, one might adopt OpenAI's recent characterisation of AGI, seen as an AI with the capacity to perform the work of a whole organisation. OpenAI's five-step framework implies that reaching this level requires the realisation of many specific capabilities in AI systems, including reasoning, autonomy, and innovation.

- **Artificial superintelligence (ASI).** Stronger than most operationalisations of AGI: "An intellect that is *much smarter than the best human brains* in practically every field." (Nick Bostrom, emphasis mine.)
- **Human-level machine intelligence (HLMI).** Sometimes defined as: AI systems which are able to perform a very large proportion (e.g. 90%, as in Zhang et al. survey definition) of economically relevant human tasks at least as well as the average human.
- **'AI scientist'.** Some TAI predictors narrow their focus to capabilities that would be sufficient for automating *specific* economically relevant fields, rather than a broad range of them.[9] For example, AIs that could replace human scientists and therefore automate scientific R&D could plausibly be transformative.[10]
- **A network of highly intelligent narrow systems.** See, for example, Eric Drexler's characterisations of Comprehensive AI Services and the Open Agency Model, as well as David Dalrymple's proposal for an Open Agency Architecture. The narrow systems comprising such a network could individually perform even narrower sets of tasks than 'AI scientists', but be at least as skilled as humans within their limited domains. While each component may not possess TAI-level capabilities on its own, the whole network in sum may amount to a transformative system (indeed, it could meet one of the other definitions on this list).

I consider all of the framings above as at least *plausibly* characterising a transformative system. So in the following discussion, when I talk about a 'short timeline to TAI', I have in mind any scenario in which we reach one of these capability levels, or something similar, by 2035. (And, as we'll see in the final chapter of this report, scenarios featuring different routes of AI progress might yield quite different forms of TAI.)

Of course, some of these benchmarks are higher than others. When I talk about 'timelines to TAI', I'm concerned with the arrival of the *first* genuinely transformative systems. If you believe that one of the weaker characterisations on this list (e.g. the 'AI scientist' frame) is achievable and genuinely

[8] Some argue that a system being truly 'general' isn't just about performing well at a wide range of tasks, as in the DeepMind definition, but more fundamentally about possessing an ability to generalise – that is, to apply reasoning or learnings from previously seen problems to tackle new situations. Thinking about 'artificial general intelligence' or TAI in this way has consequences for the arguments ahead, as we'll see in Chapter 1.

[9] Note that even if such systems have only a narrow range of tasks to perform, they might still need to possess high levels of 'generality', as in the previous footnote.

[10] See e.g. Holden Karnofsky's 'PASTA' framing, and Epoch's focus on scientific automation in its Direct Approach to forecasting TAI.

transformative, then you won't expect a stronger system (e.g. superintelligence) to be the first TAI, unless you think there's going to be some jump in capabilities R&D that bypasses the weaker systems altogether. The main consequence of this in the present report is that the period of 'takeoff' to superintelligence from systems that are already transformative is not what's under consideration here.

What I *wouldn't* count as TAI. Some near-future AI systems might be *indirectly* capable of transforming society, in the sense that they enable the development of more advanced AIs which do, in fact, transform society. For example, as we'll see in Chapter 2, an AI system that is skilled at all or some aspects of AI R&D tasks might set off a chain of improvements in the field that eventually results in the development of AIs with transformative capabilities – without *itself* possessing these transformative capabilities. However, I reserve the term 'TAI' for systems which are *directly* and *readily* capable of large-scale societal transformation, once implemented.

What drives AI capabilities progress?

To understand how TAI could emerge in the next ten years, we first need to understand how progress in AI capabilities actually *happens*, in general.

A simple model

On a simple model, which has sometimes been called the 'AI Triad', three inputs to machine learning systems drive progress in AI capabilities: **compute**[11], **data**, and **algorithms**. Improvements can be made on any of these inputs: for example, through increasing the quantity of compute used to train or run an AI system, the quality/quantity of its training data, or the efficiency/adaptability/scalability of its algorithms.[12] Improvements of any of these kinds typically lead to improvements in the performance of AI systems.

Most arguments for short TAI timelines boil down to the following: *within AI R&D, there are fast advancements in some combination of these three inputs, leading to fast advancements in the capabilities of frontier AI systems*.

Of course, AI progress is also shaped by external factors – there are many economic, social, and political influences at play here – but such influences can be seen as affecting AI capabilities progress insofar as they affect the rate of improvement in one or more elements of the Triad.

Where people disagree

If you accept this model, a few things are still up for debate, for example:

a. The *relative importance* of each input to the speed of capabilities improvements,

b. The *precise relationship*, in each case, between input improvements and capabilities improvements,

[11] This is often interchangeably called 'computational power' or 'computing power'. I typically use the term 'compute' in this report.

[12] Note that there are certain dependencies between the improvements made to these three inputs, some of which will come out in later chapters. It's not typically a case of improving one in isolation.

c. The *ease or likelihood* of making improvements on each input.

As we'll soon see, differing assumptions on these points can yield very different stories about AI progress and TAI timelines.

It's worth noting that the values of (a)-(c) are not static. Over the next ten years, things might change substantially if, for example:

- The AI R&D field begins to run out of certain computational resources;
- One of the three inputs approaches a fundamental limit of some kind; or
- We hit a threshold at which some aspects of AI capabilities progress can be automated.[13]

To understand and evaluate the arguments behind short timelines, it will be necessary to keep in mind a moving picture of progress in compute, data, and algorithms, spanning the next ten years.

How could AI capabilities progress be very fast?

How could AI progress be fast enough for a short timeline? This question breaks down into (at least) two distinct questions:

- Through which mechanisms could AI capabilities develop very quickly?
- What could the shape of AI capabilities progress look like?

Each question has a complex body of literature seeking to address it. Rather than providing an exhaustive overview of the relevant debates, I will focus on a few of the most common stories told by those who expect short timelines to TAI.

Key mechanisms for fast capabilities improvements

Prevalent stories

In the recent literature, the mechanisms that have *most commonly* been argued to enable fast AI capabilities improvements are:

(1) Compute scaling. Many stories emphasise one component of the Triad, compute, as the main bottleneck for AI progress. Proponents of *compute-centric models* of AI progress believe that, primarily through increasing the compute used in training[14], the current paradigm will be able to scale to arbitrarily capable systems; capabilities improvements on the pathway to TAI will not get bottlenecked on other components of the Triad. This implies a short timeline if AI systems can be fed with enough compute to constitute TAI by 2035.

(2) Recursive improvement. This is not a single mechanism for AI progress but a broad category of them. I use the term 'recursive improvement' to capture any positive feedback loops through which there are repeated

[13] Another key example is a scenario in which there are significant policy changes around AI R&D (for example, a decision by major actors to pause or slow down AI development). I largely bracket such scenarios in this report, and assume business-as-usual on the side of AI governance.

[14] Very recent evidence (e.g. from OpenAI's o1 model) suggests that AI systems also become increasingly capable with increased run-time compute. In light of this, 'run-time compute scaling' is increasingly featuring in stories of future AI capabilities progress, and is discussed in Chapter 1 of this report. However, I focus primarily on the prospect of scaling training compute here, for reasons I elaborate on in Chapter 1.

improvements to the ability to improve AI systems.

Although this definition includes, say, the investment feedback loops that are currently supporting AI development, other types of recursive improvement have the potential to have even *stronger* effects on AI capabilities progress. Indeed, in the context of short timelines, the most popular recursive improvement stories posit the future emergence of feedback loops under which AI systems *themselves* drive capabilities improvements by making contributions to AI R&D, and get better at doing so with each corresponding improvement. These 'direct' feedback loops could have stronger effects than the more 'indirect' forms of recursive improvement we're already seeing.[15]

This implies a short timeline if AI systems can be deployed to make sufficiently fast improvements to AI capabilities within the next decade.

These mechanisms are not mutually exclusive, but could in fact be complementary drivers of fast AI progress. Indeed, many who expect short timelines to TAI believe that capabilities will develop through a combination of the above drivers.[16]

Other stories

Not all stories of fast AI capabilities progress rely significantly on compute scaling or 'direct' recursive improvement. Indeed, there are some experts who maintain that it's likely that TAI will arrive by 2035, but do not think that the mainstream stories in the literature, based on these two mechanisms, represent plausible routes of getting there. People in this group often favour stories of TAI quickly arriving once some *change in approach* or '*trick*' to AI development has been deployed. Although I do outline some scenarios based on these alternative views in Chapter 3, my overall focus for the rest of this report will be on the key mechanisms described above, due to their prevalence in the literature.

Key shapes of fast capabilities progress

As well as the *mechanism* for capabilities improvements, the *pattern* in which those improvements occur also affects the timeline to TAI. Put differently, arguments about the timeline to TAI cannot be cleanly separated from arguments about the trajectory to TAI.

Possible shapes of the trajectory to TAI that could support a short timeline include:

- A major discontinuous jump in capabilities suddenly bringing about TAI at some point in the next ten years.
- A continuous pattern of capabilities improvements which features no major 'discontinuity', but is fast enough (and starts soon enough) to reach transformative levels within the next ten years.

[15] Note that the capabilities threshold at which this 'direct' form of recursive improvement can begin could at least *plausibly* be reached before systems we would call TAI have been developed (i.e., it's *plausibly* a pre-transformative threshold). The idea here is that we might get AI systems which can have major impacts on the field of AI R&D before we get AI systems which are capable of more *broadly* transforming the world – and that deploying systems of the former kind would influence the rate of progress towards the latter. The arguments of Chapter 2 proceed based on this assumption, though it is not certain.

[16] In fact, the most common story of progress from today's AI systems to superintelligence involves a combination of compute scaling (up to AGI or similar) and recursive improvement (from there onwards). This particular story will not be a focal narrative within this report, since it relegates the key recursive improvement dynamics to the *post-AGI* period, occurring *beyond* the likely point at which first systems with genuinely transformative capabilities have been developed.

Both of these patterns could be consistent with many different mechanisms of fast AI capabilities progress, including those characterised earlier in this section. However, some combinations of mechanism and pattern may be more likely than others: for example, some have argued that a discontinuous jump in capabilities is most likely to be achieved through extreme recursive improvement dynamics.

Since the relationship between timelines and patterns of capabilities progress is so complex, Convergence's *AI Clarity* team may tackle this directly in a separate research piece.

The present report will primarily focus on the identified *mechanisms* for fast capabilities progress. However, as noted, it will not be possible – or desirable – to completely extricate this from considerations around the trajectory of this progress. (For example, it will become necessary to briefly consider and compare different trajectories of capabilities progress when examining the potential effects of recursive improvement in AI R&D).

A roadmap for the report

Over Chapters 1 and 2, I explore the compute scaling and recursive improvement mechanisms in turn. For each mechanism, I consider:

- How it could drive AI progress and result in TAI by 2035.
- How it could fail to produce TAI by 2035.

In doing so, I will evaluate some of the most prominent arguments around compute scaling and recursive improvement in the literature, and identify key points of disagreement. These chapters will illuminate compelling stories in which short timelines appear to be likely, but which face many uncertainties and areas of serious contention. Still, no refutations appear to be decisive against the plausibility of short timelines, even if they cast some doubt over their likelihood.

Chapter 3 aims to characterise, and provide more concrete descriptions of, a set of short timeline scenarios in which one or a combination of these two mechanisms drive AI capabilities progress. I also motivate and outline a few notable short timeline scenarios which *do not* rest heavily on either compute scaling or recursive improvement.

In the Conclusion, I'll reflect on the findings of these chapters. Through these reflections, I will argue for the plausibility of short TAI timelines on the following grounds:

- There are several different routes through which TAI could conceivably be achieved in the next ten years. If AI capabilities progress is slow or begins to plateau on one route, other mechanisms could soon kick in through which TAI might still quickly be achieved.
- Short timeline scenarios are compatible with a variety of different background assumptions (for example, about scaling, the current

paradigm, and the strength of different drivers and restraints of AI capabilities progress).

- The body of evidence in support of short timeline scenarios is rapidly growing. With new developments in the AI field, the state of this debate seems to be shifting more and more towards short timelines as a likely outcome of capabilities R&D efforts.

Chapter 1: Compute scaling

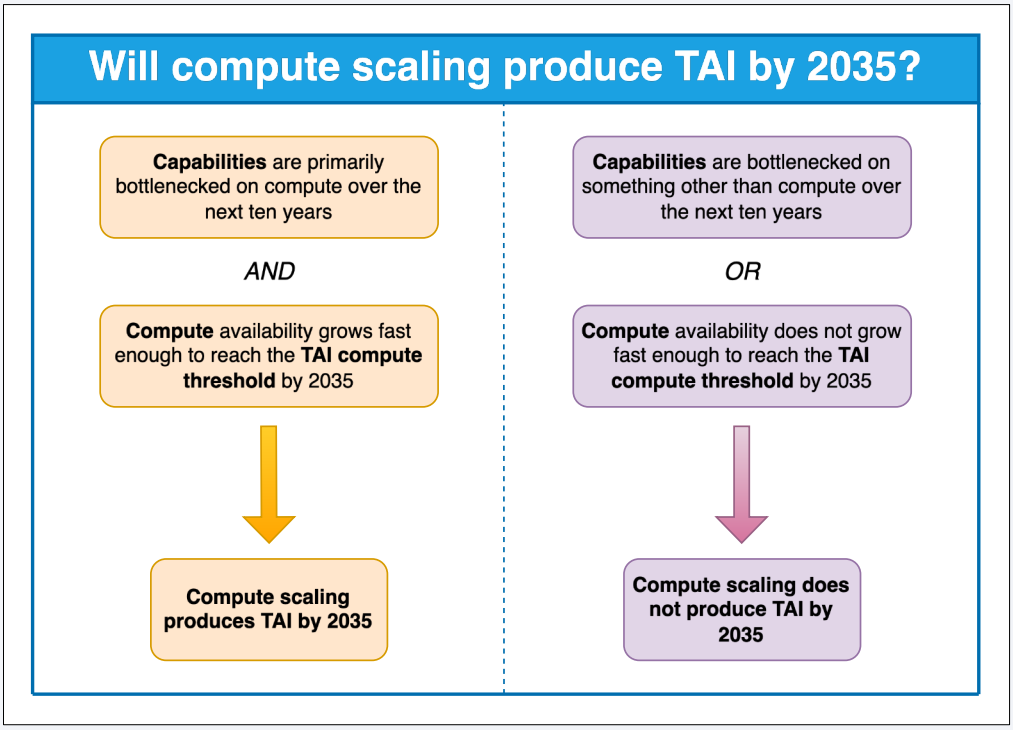


Figure 1.1: Will compute scaling produce TAI by 2035? A high-level breakdown of this chapter's debates.

What is compute scaling, and how could it produce a short TAI timeline?

The scaling hypothesis and compute

Arguments that a short timeline will be achieved through compute scaling typically rest on the *scaling hypothesis*. This is the assumption that progress towards TAI can be achieved largely by scaling up the current paradigm (i.e. neural networks with deep learning algorithms). As characterised by Gwern:

> "The scaling hypothesis regards the blessings of scale as the secret of AGI: intelligence is 'just' simple neural units & learning algorithms applied to diverse experiences at a (currently) unreachable scale. As increasing computational resources permit running such algorithms at the necessary scale, the neural networks will get ever more intelligent."

That is, some believe that 'scale' (achieved by "increasing computational resources") is and will remain the main driver of progress towards TAI over the coming years.[17] But what does this mean, exactly?

Looking at this in terms of the Triad of compute, data, and algorithms, a proponent of the scaling hypothesis might argue as follows:

[17] It's not clear that this is precisely Gwern's *own* view. As noted in the previous chapter, the dynamics of *what is driving progress* and *how* might change substantially if, for example, we reach a capabilities threshold at which AI capabilities progress can largely be automated. Many people who adopt the scaling hypothesis, possibly including Gwern, do think something like this will happen in the next ten years (but that crucially, before that point, progress will be driven by scale).

The present chapter is specifically focused on examining arguments about scale as a mechanism for AI capabilities progress. I do not explicitly consider how AI R&D automation could change the dynamics of progress over the next decade until Chapter 2.

COMPUTE, DATA, ALGORITHMS – ACCORDING TO PROPONENTS OF THE SCALING HYPOTHESIS

"Improvements on all three components of the Triad will be necessary for sustaining AI progress over the next decade. However, TAI can be achieved with systems belonging to the existing paradigm, provided that the 'computational resources' – i.e. the **compute** and **data** – used to run these deep learning algorithms are adequately scaled up. To the extent that **algorithmic improvements** are necessary for developing TAI, these improvements will heavily depend upon increased access to compute and data (e.g. for running enough experiments). In that sense, achieving TAI will *primarily* be a matter of scaling up compute and data in existing systems."

What is the relative importance of compute vs data, then, on the pathway to TAI?

The first thing to note here is that, due to dependencies between compute and data as inputs to AI systems, capabilities progress is not a case of just improving one input in isolation from the other. For example, increasing the quantity of data used by AI systems is required for making the best use of increased compute. (Shortly, once the idea of 'scaling laws' has been introduced, I'll make note of a more precise relationship that has been observed between these inputs in the case of LLMs.)

Opinions vary on what this will mean for the comparative roles of compute and data in developing TAI. I'm most interested in the views that are common amongst proponents of the scaling hypothesis *who believe that short TAI timelines are likely*. People in this subgroup have typically adopted a **compute-centric variant of the scaling hypothesis**, under which AI capabilities progress won't get significantly bottlenecked on the quantity or quality of available data within the next ten years, and as a result, the amount of compute used to train frontier models is what will largely determine capabilities progress over this time period.[18] I use the term 'compute scaling' to refer to AI capabilities progress driven by increased compute in this way, in line with these compute-centric variants of the scaling hypothesis.

Compute scaling views complement Richard Sutton's 'bitter lesson' that human ingenuity is not what's most important in AI R&D; at the end of the day, progress is mainly a matter of leveraging compute.

What evidence is there for the scaling hypothesis?

So far, progress in neural network capabilities has been in line with the scaling hypothesis: increased compute has closely corresponded with capabilities improvements. For example, GPT-3 massively outperformed GPT-2 with roughly the same architecture but more training compute. Some experts were

[18] If you expect data to soon overtake compute as the main bottleneck for AI progress, then you aren't very likely to endorse a short timeline (unless you appeal to other key drivers of progress, such as recursive improvement, or some innovative change in approach to AI development). This will become clear when we consider 'data as a bottleneck' later in this chapter.

surprised by this. Indeed, as Gwern noted in 2020:

> "GPT-3 [is] the largest neural network ever trained, by over an order of magnitude… To the surprise of most (including myself), this vast increase in size did not run into diminishing or negative returns, as many expected, but the benefits of scale continued"

Increased training compute also seems to have underpinned a major jump in capabilities from GPT-3 to GPT-4.[19]

Many experts think that this trend will continue over the coming years. In fact, some expect neural network capabilities to increase in a predictable way as training compute increases: in recent years, there have been some notable efforts to formalise this relationship between scale and performance in the form of empirical scaling laws for LLMs. For example, the scaling relationships observed by OpenAI (2020) and DeepMind (2022) are visualised below.

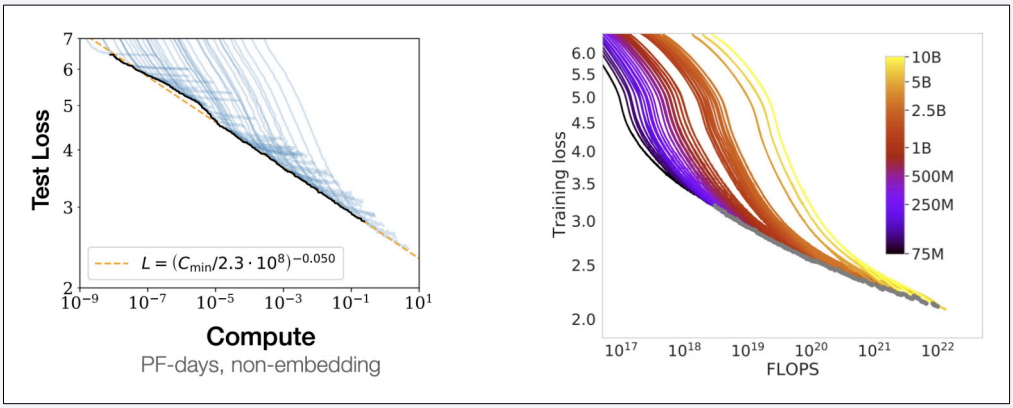


Figure 1.2: Empirical scaling laws. Relationships between compute and test loss (left) / training loss (right) in AI models observed by OpenAI (left) and DeepMind (right). Compute is measured in peta-FLOP days on the left, and in FLOP on the right. In both cases, as compute levels are increased, loss diminishes (i.e., model performance improves). Note that DeepMind's colour coding is according to the number of parameters of the model. Sources: OpenAI (2020) and DeepMind (2022).

NOTE ON THE DEEPMIND RESULTS AND 'CHINCHILLA OPTIMAL SCALING'

The scaling laws proposed by DeepMind are often called 'Chinchilla scaling laws' since they are based on experiments with a model called Chinchilla.

According to DeepMind's findings, the most efficient scaling of neural networks is achieved through scaling model parameters and training tokens in roughly equal proportion (such that the dataset is scaled in proportion to the square root of the total amount of training compute).

Elsewhere in this section, I use the term 'Chinchilla-optimal scaling' to capture any scaling in accordance with this optimal ratio.

[19] For a visualisation of the progression of GPT capabilities against (effective) compute since 2018, see the graph titled 'Base Scaleup of Effective Compute' in Chapter II of Aschenbrenner's *Situational Awareness*. Here, Aschenbrenner describes a capabilities jump from GPT-2 to GPT-3 as akin to that of a preschooler to an elementary schooler; and from GPT-3 to GPT-4 as that of an elementary schooler to a smart high schooler.

Extrapolating from this, one might expect that with enough compute, LLMs will eventually scale all the way to transformative-level capabilities. (Shortly, I'll discuss the further assumptions required for compute scaling to result in TAI *by 2035*.)

However, there are some reasons to believe the scaling trends seen thus far will break down at some point in the near future, as I'll highlight in 'Capabilities progress is bottlenecked on something other than compute'. And in fact, some very recent developments might be taken to suggest that the performance gains from increased training compute in LLMs are *already* plateauing: as of November 2024, there are reports that OpenAI's unreleased GPT-5 model is only a modest improvement on GPT-4, representing a smaller leap forward in AI capabilities than its predecessor models.[20] It is not yet clear whether we are actually witnessing diminishing returns from compute scaling, or just a temporary blip in the trajectory of GPT progress.[21]

Even if this means that AI developers won't get (in the words of a recent Forbes article) "the same amount of juice" from increases in training compute as they did previously, a version of the compute-centric scaling hypothesis may still hold. For example, there may be another axis of compute scaling for developers to exploit in the coming years: much like training compute, the compute used to *run* an AI system can also be increased. (I call this 'run-time' compute, but elsewhere it has been referred to as 'inference' or 'test-time' compute.)

Below, I discuss the prospect of scaling up run-time compute, due to the increasing relevance of this AI development strategy to the timelines debate. However, it won't constitute a major argumentative thread for this report; in fact, this report was already substantially written before much of the relevant evidence on this subject was even published.

Scaling training compute or scaling run-time compute?

The discussion so far in this chapter has concerned performance improvements from increased training compute. But there's also some evidence of AI system performance improving with run-time compute. OpenAI's September 2024 report on its o1 model, which uses chain-of-thought reasoning to solve complex problems, has provided notable support for this phenomenon. For example, the model's accuracy on a prestigious mathematics exam (AIME) was seen to increase smoothly as its run-time compute (called 'test-time compute' in the figure below) was increased, given a fixed amount of training compute (called 'train-time compute' in the figure below):

[20] For further discussion, see e.g. TechCrunch, Forbes, and Erik Hoel.

[21] As noted in a recent Forbes article on this subject, "a temporary reduction doesn't mean that the line [of scaling-based capabilities improvements] is going to veer off into a plateau for any significant length of time".

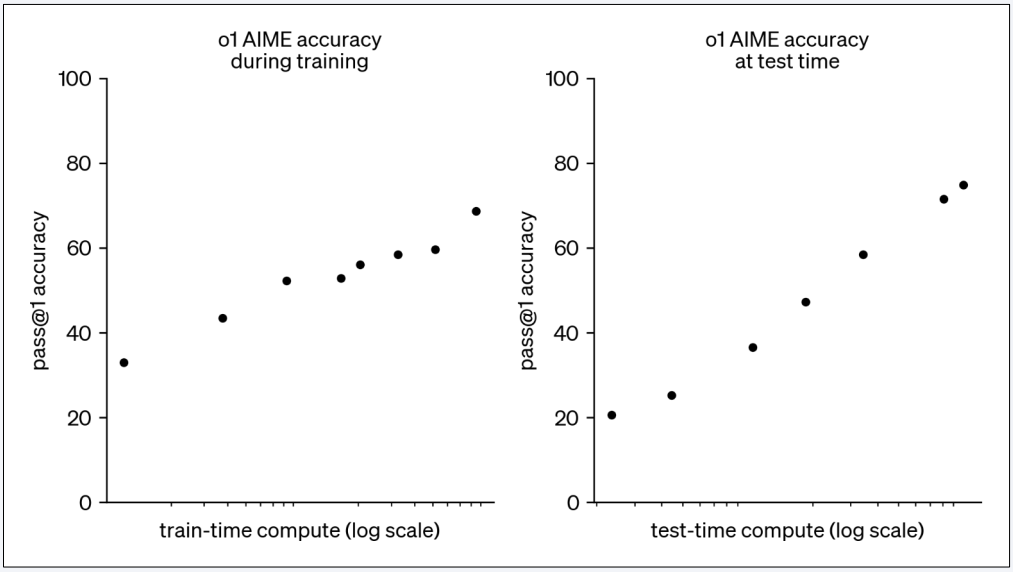


Figure 1.3: Scaling comparison, training vs run-time. Left: performance of o1 model as train-time compute is increased. Right: performance of o1 model as test-time compute is increased. *Source: OpenAI*.

Even more recently, o3 – a successor to the o1 model – has provided further evidence in the same direction.

These are significant results for debates over AI progress in general, and the conversation about scaling in particular.

According to ML researcher Nathan Lambert, this data "confirms the existence of new scaling laws – inference [i.e. run-time] scaling laws". If correct, this observation points to another dimension on which AI capabilities could be improved. And run-time compute could be an extremely powerful dimension of progress, because:

- **It offers a new route of compute scaling if the improvements from increased training compute plateau.** If the alleged slowdown in progress from GPT-4 to GPT-5 is to be interpreted as a sign that improvements from scaling training compute are breaking down, this doesn't mean that TAI cannot emerge quickly through *any* form of compute scaling. We could still see a short timeline to TAI, driven (in some part) by increased run-time compute in AI systems.
- **The possibility of exploiting an overhang from training.** An AI system, once trained, typically uses much less compute to run on than it required in training. Building this into our stories of AI progress, *alongside* the prospect of scaling training compute, will likely point towards even shorter timelines to TAI. However, it's not clear whether labs will be able to effectively stack the benefits of scaling training compute with the benefits of scaling run-time compute, or if practical constraints (such as hardware specialisation or the lab's competing resource demands) will force difficult trade-offs between the two.

The prospects for performance improvements through increased run-time compute have been mentioned in the literature for some time. For example, in

his 2023 report on takeoff speeds, Tom Davidson illustrates how trading off training compute for increased run-time compute shortens timelines according to his model.[22]

In light of OpenAI's recent work, 'run-time compute scaling' is now increasing in prominence within arguments for short TAI timelines. Historically, though, run-time compute has not been afforded a comparably sized role to training compute in stories of fast AI progress via compute scaling. Moreover, the implications of results like those very recently published by OpenAI are not yet fully understood.

Given this, I have explicitly framed the core arguments of this chapter in terms of scaling training compute (though I do briefly mention the prospect of increasing run-time compute to help overcome some specific challenges; see 'The current paradigm as a bottleneck'). However, much of what follows would also be applicable in the case of scaling run-time compute (whether done in isolation from, or alongside, increases to training compute). Moreover, the reader should bear in mind throughout this chapter that the potential for developers to adopt alternative compute scaling strategies strengthens the argument for expecting short TAI timelines, and provides a potential line of response to some of the sceptic's objections.[23]

The route to transformative AI

If increasing compute *does* continue to drive increased capabilities in neural networks over the next ten years (i.e. if scaling laws continue to hold), this might result in the emergence of TAI-level capabilities by 2035. Of course, this still depends on (i) the speed of compute growth and (ii) the amount of compute that is required to train TAI. In this subsection, I discuss these two additional variables in turn.

The key claims which must hold true for compute scaling to result in TAI by 2035 (and their relationships to one another) may be represented as follows.

[22] See also: Aidan McLaughlin's 'AI Search: The Bitter-er Lesson' for some interesting research on the role of increasing run-time compute in AI capabilities progress. This also predates OpenAI's publication of o1's performance results.

[23] In this section, I distinguished training compute from run-time compute. But in fact, training compute can be sub-divided into two further categories: compute for 'pre-training' and compute for 'post-training' (the latter phase is sometimes referred to as 'fine-tuning'). Most talk of compute scaling has concerned the compute used in the pre-training phase, but Jensen Huang has recently emphasised each of the three phases of pre-training, post-training, and run-time as important dimensions for compute scaling. (See also this quick summary of Huang's argument from Michael Dell.)

So, if performance gains from increased pre-training compute are now plateauing, there are actually *two* other dimensions of compute scaling that developers could explore: post-training and run-time. I focus on the prospect of scaling run-time compute here because it (1) has garnered more attention in recent discourse and (2) seems to be a better candidate for improving AI performance across a *range* of tasks, towards general intelligence, as opposed to the more narrow specialisation that is enabled in the post-training phase.

However, the role of post-training compute in AI development might grow in future. See, for example, Dario Amodei's thoughts on the subject in his podcast appearance with Lex Fridman.

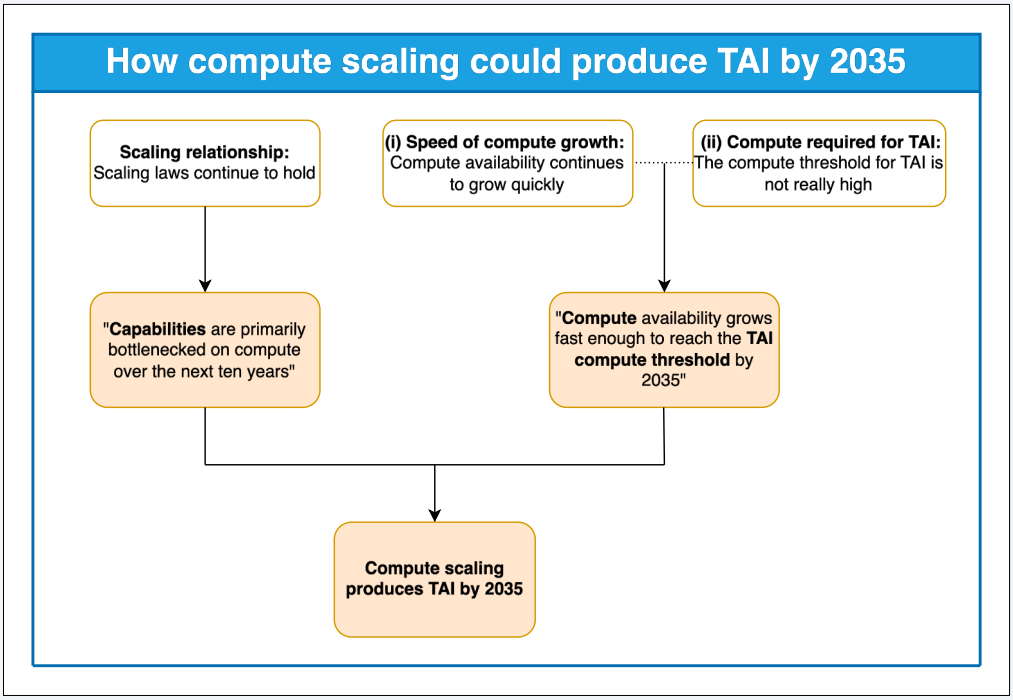


Figure 1.4: How compute scaling could produce a short TAI timeline. Previously, we discussed the 'scaling relationship' (left). In this section, I turn to discuss variables (i) 'speed of compute growth' and (ii) 'compute required for TAI'. Note that the dotted line indicates that the two claims it connects are best understood with respect to one another: what it means for a compute threshold to be 'really high' is relative to the anticipated pace of compute growth; what it means for compute to grow 'quickly' is relative to the threshold for achieving TAI-level capabilities.

How fast could compute grow over the next ten years?

With respect to variable (i) above ('speed of compute growth'), it will be illustrative to reflect on the last few decades of AI progress.

Epoch has identified 'three eras' of machine learning which have seen different rates of growth in the compute used to train frontier models.

Between the 1940s and 2010, the physical compute used to train AI systems doubled roughly *every 20 months*. The growth seen in this 'pre-deep learning era' is noted to roughly track *Moore's Law*: the historical observation that the number of transistors on a microchip has doubled approximately every two years. Increases in the compute used to train AI systems have been, to some extent, driven by these increases to microchip density.

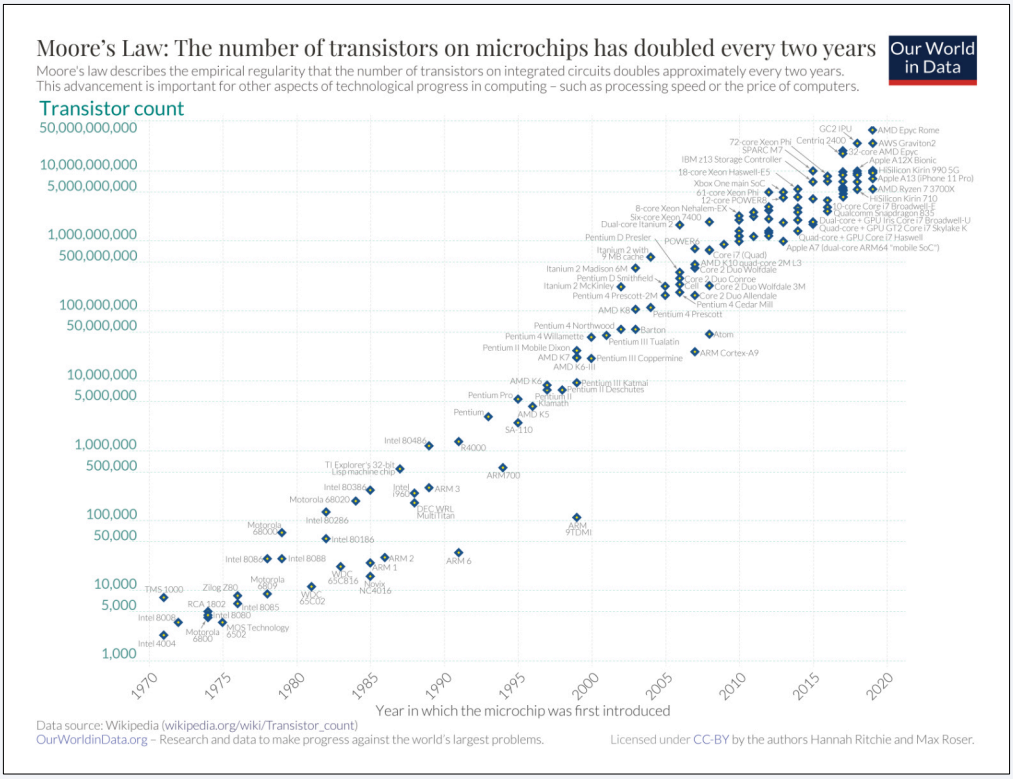


Figure 1.5. Graph illustrating Moore's Law. *Source: Max Roser, Hannah Ritchie and Edouard Mathieu (2023) – "What is Moore's Law?" Published online at OurWorldinData.org. Retrieved from: 'https://ourworldindata.org/moores-law' [Online Resource].*

Moore's Law is constrained by fundamental physical limitations and is therefore expected to break down eventually. There is some dispute over when this will happen. Some believe that Moore's Law has already ended, and some expect its physical limits will be reached before the end of the 2020s, while others are more optimistic.

Even if we are witnessing a death of Moore's Law, the recent data collected by Epoch doesn't show any signs of slowdown in compute growth. In fact, the growth rate appears to have increased since 2010:

- Somewhere between 2010-2012, we entered the 'deep learning era', marking a paradigm shift in ML. This era has seen training compute double *once every six months*.[24]
- If we treat the recent influx of large-scale models developed by big corporations as a category in its own right, it seems (speculatively) that the trend line bifurcates around 2015-2016. In the 'large-scale era' characterised by the emergence of these behemoth models, compute has been doubling *once every ten months*. (This growth rate is still twice as fast as pre-2010 growth.)

[24] In fact, more recent Epoch research suggests that this might even slightly underestimate compute growth: training compute for frontier models might in fact be increasing as much as 2.5x every 6 months.

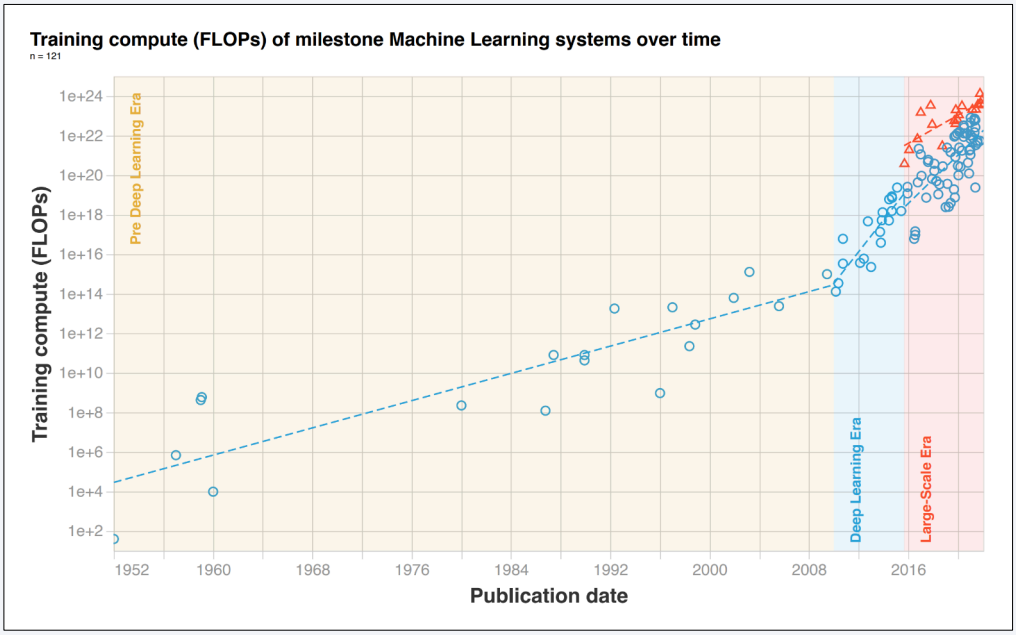


Figure 1.6: Training compute growth since the 1950s. Compute used to train 121 'milestone' ML models between 1952 and 2022. Three 'eras' are distinguished. *Source: Epoch*.

What's driving this growth, and will it continue? The amount of physical FLOP available to train a system is determined by several factors, including:

1. The number of chips
2. The density of chips (the variable that Moore's Law describes)
3. The design of the chips
4. The design of the data centre (i.e. all components of the data centre above the level of chips)

Growth in physical compute has been driven by advances on all four of these fronts (enabled by increased investment[25] and R&D effort). Importantly, AI systems today are also *able to do more* with the same amount of FLOP than earlier systems: that is, 'effective' compute is growing even faster than physical compute. This has been driven by algorithmic progress (which contributes around 3x increases to effective compute every year, according to research by Epoch), as well as improvements in the quality of training data. (From here onwards, the arguments of this chapter can largely be interpreted in reference to scaling up the 'effective' compute used in AI training runs, rather than just physical compute.)

Those who anticipate short TAI timelines via compute scaling believe that, through investment and R&D effort, advances in some or all of these areas (i.e., the four listed above, and algorithmic efficiency) will continue at pace over the next ten years. This would result in significant increases in the effective compute available to frontier systems over the next ten years, eventually reaching some threshold for compute that is sufficient for training TAI.

How much compute is needed to train TAI?

This threshold can be estimated in a few ways, including:

[25] Epoch estimates that the total spending on the final training run for a frontier model has increased 2.4x per year since 2016. This exponential growth in investment has been essential for supporting recent capabilities progress.

- Via comparisons between artificial intelligence and biological intelligence. See e.g. Ajeya Cotra's Biological Anchors report.
- Via direct appeal to empirical scaling laws. See e.g. Epoch's Direct Approach.

Often, these approaches are used to produce *probability distributions* of the effective compute required to train TAI, rather than a single threshold. Under Epoch's model, the median of the resulting distribution (visualised below) is around 10^34 FLOP.[26] Ajeya Cotra's model features six separate probability distributions for the compute requirements corresponding to each of six biological 'anchors', with medians between 10^29 and 10^41 FLOP.

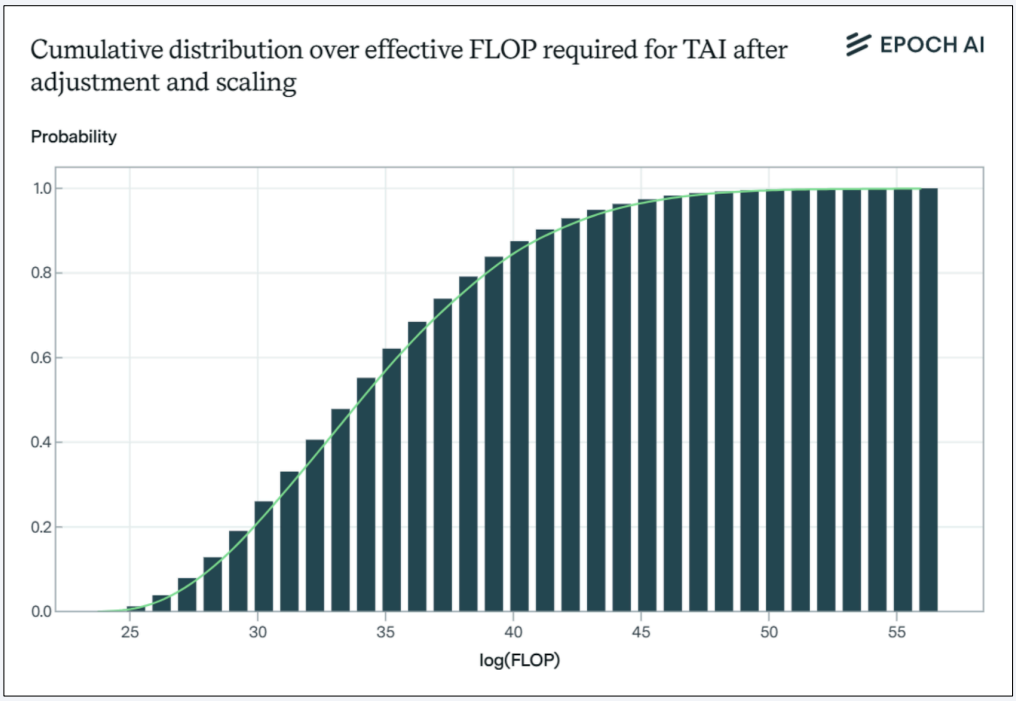


Figure 1.7: Probability distribution of the effective FLOP required to train TAI, according to Epoch's 'Direct Approach' model. *Source: Epoch*.

Both of these models are described in some detail in my previous post on Timelines to Transformative AI. More recently, I also conducted a much closer examination of Epoch's model with my colleague Elliot McKernon.

Forecasting TAI with compute-centric scaling models

By making projections of effective compute growth and establishing probability distributions for the amount of compute that would be sufficient for TAI, it is possible to quantitatively forecast the arrival of TAI. There have been several notable attempts to do this, including the examples listed in the table below.

Forecast models based on compute scaling often place significant weight on short timelines. Indeed, all three models compared below have a greater than 15% chance of TAI arriving by 2035.

[26] That is, after some adjustment to parameters in correction of an apparent mistake in the scaling law. See my earlier piece with Elliot McKernon, *Now THIS is forecasting*, for more detail on this correction. And note that other variants of the model yield slightly different medians (of 10^33 FLOP, 10^35 FLOP, or 10^36 FLOP).

Timeline predictions from a selection of compute-centric forecast models

Forecast model	Median TAI arrival date	Probability of TAI by 2035
Ajeya Cotra's *Biological Anchors* model	2052	16%
Tom Davidson's *Compute-centric takeoff speeds* model	2043	20%-40%[27]
Epoch's *Direct Approach* model	2033	50-60%

Table 1.1: Timeline predictions based on compute scaling. Note that TAI has been operationalised in different ways in all three models.

Beyond these probabilistic forecast models, there's also a wealth of personal timeline predictions by experts which have been based, wholly or partially, on the assumption of compute scaling. One notable example comes from Ray Kurzweil, whose own compute projections have led him to consistently endorse a prediction of TAI by 2029 (specifically: human-level AI by 2029) over the past several decades.

Those who endorse the scaling hypothesis need not straightforwardly accept the bottom line results of any particular model or prediction; indeed, the complex web of variables and limitations of empirical data mean there is a great deal of uncertainty around all of the results listed above. However, these examples illustrate how believing in the power of compute scaling *could* lead you to consider short TAI timelines as a serious possibility – perhaps even the median outcome of AI progress.

Why compute scaling might not produce a short TAI timeline

There are many sceptics who believe that the compute scaling pathway characterised above will fail to produce a short TAI timeline.[28]

Below, I explore several such objections. I divide these up into two broad categories, according to whether they reject or accept the compute-centric scaling hypothesis: that compute will be the primary driver of, and bottleneck for, capabilities progress on the road to TAI.

Category 1 sceptics say: "The compute-centric scaling hypothesis is false; capabilities progress will be bottlenecked on something *other* than compute, which will be harder to make improvements on".

Category 2 sceptics say: "The compute-centric scaling hypothesis may be true, but the compute threshold necessary for TAI cannot plausibly be reached by 2035 (because the growth in compute available for training runs will be slow, or because this threshold is so far away)".

[27] It seems that an exact figure is not explicitly stated in the reports or online playground. However, it *is* specified that P(TAI by 2035) is in the 20%-40% range, and from looking at a graph, it appears to be around 30%.

[28] Notable sceptics of compute scaling as a route to short TAI timelines include François Chollet (whose arguments are represented in this section) and Gary Marcus (whose arguments against the compute scaling pathway can be found here). Note, however, that scepticism of short timelines *via compute* scaling doesn't necessarily mean scepticism of short timelines *in general*. For example, although Chollet doubts that TAI can be achieved from scaling up the current paradigm, he does still take seriously the prospect of short timelines (as will become evident in Chapter 3). Still, many people who adopt views of this kind are more generally sceptical of arguments for short timelines.

This is not intended to carve out a clean division of the landscape of counterarguments. The categories I'm gesturing towards could be variously interpreted, and the membership of a counterargument to one category over another is sometimes a matter of framing rather than a matter of principle.

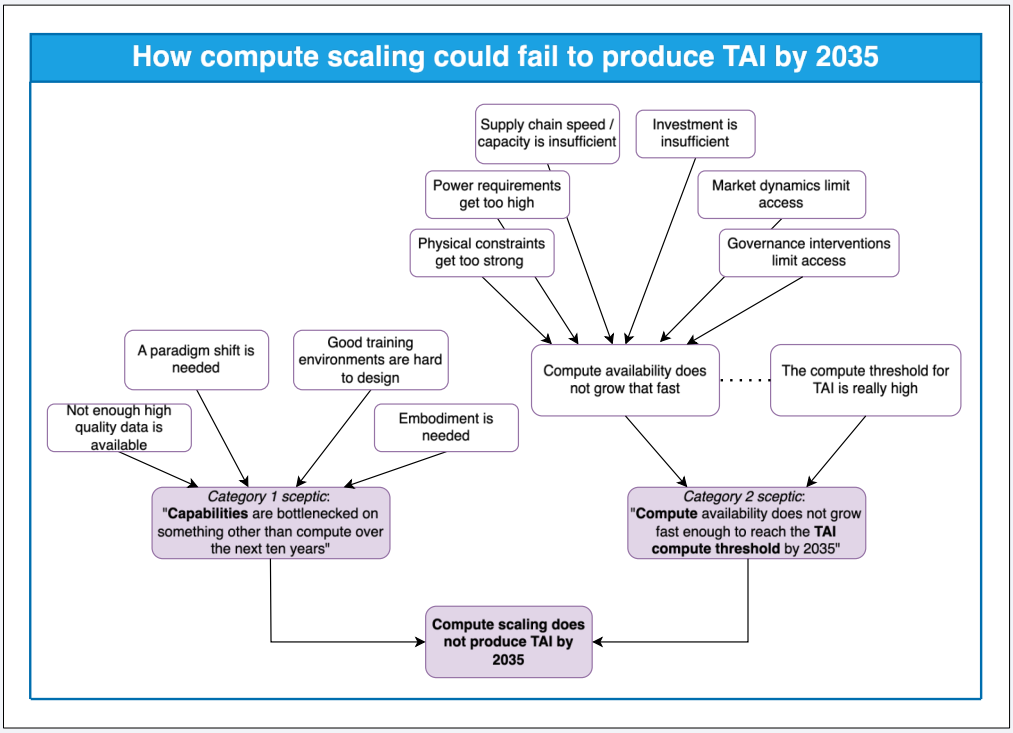


Figure 1.8: How compute scaling could fail to produce a short TAI timeline. This diagram represents the overall structure of the arguments below. As in Figure 1.4, the dotted line indicates that the two claims it connects are best thought of in relation to one another.

Category 1 sceptical arguments: Capabilities progress is bottlenecked on something other than compute

Sceptics often claim that those who believe compute scaling will soon produce TAI have overemphasised the importance of compute in AI capabilities progress. As such, they have failed to account for other important bottlenecks.

Potential bottlenecks to AI progress include the two *other* components of the AI Triad (data and algorithms), as well as certain factors external to the Triad model which are considered to be important under competing views of AI progress. I explore several possibilities below.

Data as a bottleneck

Training bigger and better AI models requires bigger and better datasets. This means that even if developers can access enough compute for a TAI-scale training run, they cannot use this compute to train a TAI-scale model unless datasets are also scaled up to an appropriate degree, and are of a high enough quality. At some point in the next decade, requirements on the quantity or quality of data used in training might therefore overtake compute requirements as the primary bottleneck for capabilities progress.

How difficult will it be to meet data requirements?

One question here is: how much data will be needed to train the first TAI? This is not clear cut, but in Will scaling work?, Dwarkesh Patel's imagined sceptic claims that *efficiently* scaling current models up to the FLOP benchmark Epoch has estimated for TAI-level capabilities[29] will require around five orders of magnitude more data than is currently available to developers. It might be possible to achieve TAI-scale training runs with somewhat less data than this if developers do not follow Chinchilla-optimal scaling[30] – "but", Patel's sceptic says, "this can help you make up for a slight data deficit, not a 5 OOM shortfall". In other words, there's no getting around it: training TAI will require *more data than we currently have*, and probably several orders of magnitude more.

However, it seems that high quality data is in short supply. For example, according to projections by Epoch made in 2022, we have already nearly exhausted our stock of high quality language data, and will even run out of *low quality* language data somewhere between 2030-2050.[31] Access to data is also likely to become more expensive (and perhaps more tightly controlled) in the coming years, as concerns grow over data collection practices and fair compensation. This could be a serious bottleneck for capabilities improvement over the next ten years.

Responses to the data objection

Those who believe in short timelines through compute scaling have a few ways of responding to this objection.

For example, they can point to efforts in data cleaning and filtering as ways of improving the quality of available datasets. According to recent research, this could enable higher performance in models trained on much smaller datasets. However, it is not clear how many orders of magnitude of improvements this could translate into; it seems unlikely to be enough to unlock TAI-level capabilities from existing training datasets.

They can also point to the potential for using more multimodal data to train systems. Indeed, Epoch has recently updated its 2022 data scarcity projections (which were cited earlier) by factoring in the use of multimodal data, and the updated results are optimistic: the authors foresee "multimodal learning from image, video and audio data… plausibly tripling the data available for training", and consequently, don't expect data to become a significant bottleneck for AI progress by 2030. Specifically, they "estimate the equivalent of 400 trillion to 20 quadrillion tokens available for training by 2030, allowing for 6e28 to 2e32 FLOP [Chinchilla-optimal] training runs".

However, the multimodal data strategy faces a similar complaint to that of data cleaning and filtering: it's not clear exactly *how far* it can take us. By 2035, will this provide enough data to support a Chinchilla-optimal training run to the tune of something like 10^35 FLOP (which may be what's required for TAI)? In addition, it's very uncertain to what extent patterns learned from one modality

[29] Here, the sceptic selects the 10^35 FLOP benchmark noted by Epoch here, rather than (for example) the parameter-adjusted and recalculated 10^34 FLOP benchmark I cited earlier. Although this does affect the estimate for corresponding data requirements, it has no tangible impact on the *overall* argument here.
Also note that 'TAI' has been operationalised by Epoch as an AI system able to automate scientific research and development. (Specifically, Epoch selects the ability *to produce scientific papers indistinguishable from those produced by human scientists* as the performance benchmark here.)

[30] Recall that the Chinchilla scaling law (DeepMind, 2022) suggests that compute-optimal scaling is achieved through model parameters and number of training tokens being scaled in roughly equal proportions. Scaling up systems using less data and more parameters is possible, but would not be considered optimal from the perspective of the Chinchilla scaling law.

[31] However, as I'll note shortly, Epoch's more recent projections on data (which factor in the prospect of using multimodal data sources to train future AI systems) are more optimistic.

of data (e.g. images) will actually transfer into enhanced performance on another (e.g. text).

Another popular response to the data objection invokes the possibility of *creating new data*. It is often argued that with near-term increases in compute, AI systems will soon be empowered to generate synthetic datasets, e.g. through self-play, and thereby enable the gains from compute scaling to continue.[32] Effectively, increased compute in the near-term could provide a means to avoid running out of high quality data for training, thereby sidestepping the data problem.

Although Epoch researchers haven't factored the potential for synthetic data generation into their recent projections, they do "expect synthetic data to likely be useful for overcoming data bottlenecks". This strategy seems promising even if the gains from data cleaning/filtering and using multimodal data aren't sufficient over the next ten years to make up for the TAI data shortfall. As such, the success of synthetic data generation is viewed by some (e.g. Dwarkesh Patel[33]) to be a major crux of the debate around timelines. If it works really well, then compute scaling can continue to yield performance improvements unconstrained by data issues[34]; but if not, there's a strong chance that something other than scale will be needed to get us to TAI in the next decade. I therefore discuss the prospects for synthetic data generation in a little more detail below.

Will synthetic data generation work?

There's some evidence for synthetic data generation helping to improve the performance of AI systems, with the training of AlphaGo Zero as one notable example. But can this method be taken far enough to enable the training of a TAI-scale system, making up for a several OOM data deficit?

On the one hand, there are significant challenges here. Successful self-play of this scale calls for *huge* amounts of additional compute. In some domains, it might also require addressing the difficult task of getting current systems to evaluate their own performance.[35] And even if developers can meet those challenges, the applications of synthetic data may be disappointingly limited. Here, Narayanan and Kapoor argue that "synthetic data is not magic": since LLMs would be generating such datasets based on their own training distributions, they might not capture any meaningfully *new* problems. This means the benefits of synthetic data may primarily be in improving performance within narrow domains, rather than actually *replacing* (or genuinely expanding) current sources of training data.

On the other hand, it's worth noting that the researchers who are actually working at top AI labs (such as Dario Amodei) don't seem, in general, to be worried about data bottlenecking their plans for developing AGI. They haven't divulged much about *why*, but have gestured to an expectation that they'll get synthetic data to work. We can, of course, speculate about biases that might be informing these views. But perhaps we should take note of the confidence with which industry insiders – who have access to information we do not – believe

[32] AIs engaging in synthetic data generation/self-play might also introduce some positive feedback loops in the field of AI R&D, as I will discuss in the following chapter on 'recursive improvement'. Notably, AI systems which engage in synthetic data generation/self play are not sufficient *on their own* to bring about a period of what I call 'direct recursive improvement', under my specific characterisation of this term – but could do so if acting in conjunction with other AI systems which are used to automate other aspects of AI R&D. I'll return to this idea in Chapter 2.

[33] In his own words, "if some kind of self-play/synthetic data doesn't work, you're absolutely f***d – there's no other way around the data bottleneck".

[34] Assuming, of course, that there is no other major bottleneck at play.

[35] But formal verification can be used for tasks in other domains e.g. coding and mathematics.

they can overcome these supposed data issues.

Further reading:

- *Will we run out of ML data? - Epoch (and the accompanying report)*
- *Will scaling work? - Dwarkesh Patel*
- *AI scaling myths - Narayanan and Kapoor*
- *On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey - Long et al.*
- *Best practices and lessons learned on synthetic data for language models - DeepMind*
- *Can AI scaling continue through 2030? - Epoch*

The current paradigm as a bottleneck

Some objections to the compute-centric scaling hypothesis emphasise the third component of the Triad: algorithms.

The strongest objections of this kind don't just point to the need for algorithmic improvements *within current systems*, but to the need for a much bigger shift: that is, achieving TAI will require *moving away* (in some sense) from the current paradigm of neural networks with deep learning, and developing new architectures that are better equipped to handle a broad range of problems. This could take significant R&D time and effort, reducing the likelihood of a short timeline.

Such arguments are made along the following lines: although large-scale LLMs are currently performing well on many benchmarks, these benchmarks don't truly capture the capabilities or characteristics that are necessary for TAI. Moreover, due to some fundamental issues with the current paradigm, plugging additional compute into these systems is not likely to translate into high performance on the areas that actually matter. In the words of Dwarkesh Patel's sceptic, "there's no amount of jet fuel you can add to an airplane to make it reach the moon".

I'll shortly run through the details of a sceptical argument of this kind, but first: a note on the types of model it applies to, and its possibly diminishing relevance in light of very recent developments in the AI field.

A NOTE ON 'TRADITIONAL LLMS' VS NEWER RELEASES

The most influential objections of this kind have specifically highlighted the limitations of large-scale GPTs like GPT-3 and -4. As of December 2024, this discourse suddenly feels somewhat outdated, in light of the very recent releases of OpenAI's o1 and o3 models. This is especially true for o3, which came out while the present report was in its final stage of revisions, and which has apparent strengths where models like GPT-3

and -4 had apparent weaknesses.

At the current time of writing, there is not enough public information on the o-series from OpenAI to determine whether o3 should be seen as falling *within* the current paradigm, or if it in fact marks the beginning of some kind of paradigm shift. To see that there is some debate here, note that Yann LeCun has tweeted that o3 is "not an LLM (even if it uses one)".

The core argument of this section **will concern what I call 'traditional LLMs'**, referring specifically to AI models like GPT -3 and -4, Claude, and Gemini. Given the opacity around OpenAI's R&D efforts in recent years, it's difficult to pinpoint *exactly* what sets these systems apart from o3 (and perhaps some other new releases). However, I think that one key differentiator might be something like this: in what I call 'traditional LLMs', the *only substantive contributions* to model outputs come from a transformer-based neural network trained using next-token prediction.[36]

If o3 is rightly seen as an LLM, it is probably not a traditional one in this sense. François Chollet suspects that the o3 model represents "a form of *deep learning-guided program search*". That is: although it is assisted or "guided" by an underlying transformer-based neural network (specifically, a GPT), something else has been built on top of this base model to search over a space of programmes – and this 'something else' seems to be making *substantive contributions* to o3's outputs. Based on these speculations, o3 seems to represent a meaningful departure from models like GPT-3 and -4. At the very least, it doesn't seem to be recognisable as a predominantly scaled-up version of OpenAI's previous releases (i.e. something architecturally very similar, but with more training compute plugged in).[37]

Later on in this section, I'll note that o3 appears (based on the limited information we have about it so far) to be basically immune to the highlighted critiques of traditional LLMs.

If this is all correct, what does it mean for the arguments at hand?

Firstly, the overarching sceptical argument of this section may have lost some of its bite. Addressing the apparent limitations of current systems no longer seems to be quite so time-consuming and difficult to achieve, even if something like a paradigm shift is required. In fact, we *might* already be on the cusp of achieving exactly this.

Secondly, if Chollet is correct about the architecture of o3, this points to a way in which the development of TAI might end up not even involving a significant degree of further scaling: possibly, future capabilities gains could largely be based on innovations in leveraging programme search, without relying on substantial further scaling with compute.[38] (I'll pick this idea back up in 'Have we missed anything important?' in Chapter 3.) This

[36] As a result, capabilities progress in traditional LLMs (like the jumps in capabilities seen from GPT-2 to GPT-3, and from GPT-3 to GPT-4) can be seen in large part as an application of scaling up that neural network with increased training compute.

[37] I might be wrong to draw distinctions in this way. Given the lack of information about o3 and other recent OpenAI releases, I can only speculate here on the details. In fact, Chollet's above commentary on the setup of o3 is based on (educated) speculation rather than direct knowledge of the chosen architecture. However, what is evident is that there is *something* quite different about o3 which separates it from the models I class as 'traditional LLMs', even if we can't precisely locate what that *something* is. Speaking to ZDNET, Chollet said: "It is completely obvious from the latency/cost characteristics of the model that it is doing something completely different from the GPT series. It's not the same architecture, nor in fact anything remotely close."

is a point generally in favour of the believer in short timelines, but not *specifically* in favour of those who believe in short timelines *via compute scaling*.

With that said, I'll now advance the sceptical argument as it relates to *traditional* LLMs and compute-centric stories of AI progress.

What traditional LLMs have struggled to do

A feature that is common to many (but not all) conceptions of TAI is the ability of a system to generalise, i.e. to apply reasoning from previously learned problems to tackle novel situations. This feature is often called *generality*.[39] It has sometimes been distinguished from the ability of a system to simply 'memorise' problems from its training distribution.

Sceptics of the current paradigm have argued that traditional LLMs are failing to generalise. A key example of this apparent failure, cited especially by François Chollet in *On the measure of intelligence* and his recent appearance on the Dwarkesh Patel podcast, is the poor performance of traditional LLMs on Abstraction and Reasoning Corpus (ARC) problems. Although these visual puzzles are (alleged by Chollet to be) fairly simple for humans to solve, traditional LLMs like GPT-4 have struggled.

This observation matters because ARC problems are, by design, resistant to what sceptics have dubbed 'memorisation'. More specifically, since the problems are not found on the internet, solving them requires a system to apply its skills to a new situation that was not represented in its training distribution. Performance on ARC problems may therefore be seen as a good indicator of a system's level of generality.

By contrast, sceptics believe that most other commonly cited benchmarks – the ones that traditional LLMs have been doing well on – are primarily measuring the system's memorisation ability, rather than generality.

Why scaling up compute might not help us solve this problem

It's true that increasing the compute used to train LLMs has improved their performance on many tasks. But large-scale traditional LLMs, such as GPT-4, have not performed as well as one might have expected on benchmarks such as ARC, which are possibly important. Could this just be because we need *even more computational resources* funnelled into training to improve performance on these kinds of tasks, as compared to others?

Sceptics of the current paradigm, such as Chollet, don't think so. This is due to their belief that the 'successes' of large-scale traditional LLMs on many existing benchmarks have been achieved through memorisation. They argue that the training these large-scale systems have received is so extensive that solving future challenges will, in many cases, simply be a matter of repeating solutions for problems they have already seen before. By scaling a model up,

[38] Of course, programme search occurs during run-time, and therefore comes with demands on the side of run-time compute. But while deploying effective programme search will likely require devoting more compute to inference, this doesn't mean it will require significantly increasing the amount of compute used by AI systems *overall* – instead, this might just turn out to be a more efficient way of allocating the same fixed amount of compute.

[39] Terms like 'generality' are used quite variably in the literature. Here, we've defined it as the *ability to generalise* to new problems. Elsewhere, terms like 'general' or 'generally intelligent' are applied to systems which can perform well on a wide range of tasks, in contrast to those which can only do so within a narrow domain. Although these are two slightly different concepts, they are closely connected to one another: possessing strong 'generality' (i.e. a *strong ability to generalise*) would enable a narrow system to become more 'general' (in the sense of becoming competent on a broader range of tasks).

giving it more training compute and data, you enable it to memorise *even more* problems – so *of course* it will perform even better overall on tasks where it can lean on memorisation. But for tasks that require a novel insight or generalisation to a new situation, more training compute and data is just not going to help.

A slightly different way of thinking about this is that empirical scaling laws have been measuring the wrong kind of performance metric. These laws tell us how well LLMs of a certain scale can predict data from their training sets, but are silent on their performance on tasks outside of that distribution.

Why shifting away from traditional LLMs might be necessary

Taking stock here: we have a problem, and the sceptic has said that scaling training compute will not solve it. To conclude from here that some kind of paradigm shift is needed, we first need to explain why more minor adjustments on the software side, within the current paradigm of traditional LLMs, will not solve the problem. The question is: what does the sceptic think is so *fundamentally* wrong with traditional LLMs that they cannot achieve high levels of generality?

For Chollet, the key weakness of traditional LLMs here is an inherent feature of deep learning: the system can only draw on examples within its training distribution. All its 'learning', in the form of adjusting its parameters, occurs during the training phase. Once training is complete, these parameters are fixed; the system cannot update its knowledge without retraining. As a result, its ability to respond to novel problems is limited by the information it was exposed to during training.

On the other hand, the systems which actually fare well on ARC problems are more flexible in adapting to new problems, as they learn in a more dynamic way. In particular, systems based on Discrete Programme Search (DPS) have performed better on the ARC benchmark than traditional LLMs like GPT-3 and -4. In DPS, systems find ways to tackle problems by searching through a pre-defined set of primitive programmes. Crucially (and unlike systems based solely on deep learning), they continue 'learning' after training, by exploring and combining different elements from more primitive programmes within the programme space. This might be why they are (or appear to be) better suited than traditional LLMs to tackling unseen problems that require some form of novel insight.

In his podcast appearance, Chollet argues that deep learning and DPS are effectively opposites; the strengths of one approach are weaknesses of the other, and vice versa.[40] If this is the case, then although the current paradigm (of what I'm calling 'traditional LLMs') is powerful in some ways, it may not be enough on its own to achieve the capabilities or features necessary for TAI. Perhaps this means some hybrid of deep learning and DPS will be needed for continued progress in AI capabilities. As I'll note below, developers already seem to be taking steps in this direction.

[40] Discussed at timestamp 0:49:35 in his interview with Dwarkesh Patel. In fact, Chollet views them as useful for different modes of thinking: deep learning is a good fit for 'system 1' thinking, while DPS is useful for 'system 2' thinking (under Kahneman's definitions).

Implications of a paradigm shift for TAI timelines

A sceptic might argue that the argument above casts significant doubt on short timelines. If 'traditional LLMs' are inadequate for TAI, a new kind of system will need to be developed and trained – and this could take some time.

This point is gestured to in *LLM generality is a timeline crux*:

> "If LLMs are fundamentally incapable of certain kinds of reasoning, **and** scale won't solve this (at least in the next couple of orders of magnitude), **and** scaffolding doesn't adequately work around it, then we're at least one significant breakthrough away from dangerous AGI."

Until very recently, this seemed like an obvious reason to doubt short TAI timelines. However, given the latest developments in the field, it seems that *even if a shift away from traditional LLMs is needed*, this might not take much time to implement. In the language of the quote above, one might argue that we have already had, or are on the brink of, the "one significant breakthrough" that will enable the development of TAI.

Specifically, AI developers have taken what *might* be seen as a step towards realising Chollet's vision (described above) of a hybrid system based on DL and DPS, in the form of OpenAI's o3 model. Though nothing has been confirmed directly by OpenAI, the o3 model is *suspected* to utilise deep learning techniques equipped with "a new capability called 'program synthesis,' which enables it to dynamically combine things that it learned during pre-training – specific patterns, algorithms, or methods – into new configurations" (quote taken from Matt Marshall, writing for Venture Beat). In Chollet's words, "o3 represents a form of *deep learning-guided program search*".

Apparently vindicating Chollet's predictions, o3 has performed extraordinarily well on the ARC benchmark, achieving a breakthrough result of 76%; this is the first time an AI system has outperformed a human score on the test.

Even if these speculations about the architecture of o3 are correct, it's still too early to tell whether further efforts to harness programme search would yield more performance gains of the kinds relevant for developing systems with high levels of generality. As a result, the believer in short timelines cannot prove that a successful shift in direction of this kind could actually result in TAI within the next ten years. But it's worth noting that, even before news of o3 had broken, Chollet himself declared that he believes some form of TAI is "likely in the next 10-15 years" despite his scepticism of traditional LLMs and of compute scaling as a pathway for progress. Clearly, he imagines his proposed strategy for improving AI generality (or a similar one) can feasibly be implemented in the near-term, even though it will require a shift away from traditional LLMs.

If Chollet's vision of a hybrid model, as opposed to a traditional LLM, is what gets us to the first transformative systems, this might end up being less a success of the *compute scaling mechanism* that this chapter is focused on and more a success of a *new approach to AI development* (i.e., the effects of a

particular innovation). I'll return to this point in Chapter 3, where Chollet's personal view will be captured under an 'alternative scenario' of AI progress.

For now, it's worth noting that a shift away from traditional LLMs, in the direction of o3 or otherwise, may not even be *necessary* on the pathway to TAI. If so, it doesn't matter whether a paradigm shift would be achievable in the next ten years; continued progress with traditional LLMs could be enough to produce a short TAI timeline.

Why a paradigm shift may not even be necessary

Importantly, it's not clear that the 'limitations of traditional LLMs' highlighted thus far represent serious challenges. There are several reasons to think that these alleged limitations (a) could be easily and quickly addressed within-paradigm, or (b) are actually of little relevance for the realisation of TAI. I highlight five of these reasons below.

"Couldn't some form of unhobbling, such as scaffolding or tooling, significantly improve the generality of current systems?" The poor performance of current models on the ARC benchmark might not be a fundamental defect of traditional LLMs that demands a paradigm shift. Instead, we might argue that LLMs have latent capabilities which are yet to be unlocked.

Aschenbrenner argues that utilising 'unhobbling' techniques, such as scaffolding and tooling, can unlock these capabilities and significantly improve performance. According to research by Epoch, these techniques "offer [performance] gains equivalent to training with 5 to 20x more compute at less than 1% the cost".

Amongst other things, this *might* solve the apparent generality problem. Moreover, such solutions could plausibly be implemented within the next ten years – indeed, Aschenbrenner characterises them as "small algorithmic tweak[s]" – without requiring a significant shift away from traditional LLMs.

"Couldn't easy wins in scaling up run-time compute, exploiting the overhang from training, significantly improve the generality of current systems?" Perhaps just giving LLMs more time to 'think', in the form of increased run-time compute, is what's needed to improve their overall reasoning abilities. After all, in both o1 and o3, improved performance on the ARC benchmark has been partly attributed to the successes of chain-of-thought (CoT) reasoning enabled by increased run-time compute.

Of course, we've noted that o3 is (probably) not itself a traditional LLM, and it's unclear where o1 stands amongst these (very roughly defined) categories. But at least *in theory*, leveraging CoT + increased run-time compute need not involve a significant shift away from traditional LLMs: for example, a traditional LLM could generate chains of thought through next-token prediction, without leveraging programme search. And *in practice*, there's some reason to believe that the improvement on the ARC benchmark from GPT-4 to o1 was primarily due to the latter's increased run-time compute, even

if there are other very substantive differences between the two models: indeed, in response to the successes of o1, ARC prize cofounder Mike Knoop commented outright that "the big new story is test-time [i.e. run-time] scaling. We believe iterated CoT genuinely unlocks greater generalization".

Either way, a strategy emphasising run-time compute could plausibly be implemented within the next ten years, given the overhang from training that could be exploited here.

"The specific *ability to generalise* isn't that important anyway; *general intelligence* is what matters, and multi-modality in traditional LLMs is already a good demonstration of it." Traditional LLMs may not be demonstrating a strong ability to generalise to unseen tasks. However, they are succeeding, with scale, in performing an increasingly wide range of tasks. A single AI system can now process and generate text, code, images, audio, and video.

This isn't necessarily a sign of *generality*, in the sense in which the sceptic is using this term. But this new ability to perform a wide range of tasks, even if those tasks are largely within-distribution or 'memorised', is something we could label as increased *general intelligence*. And we might argue that it is general intelligence, not generality, that is actually important for the realisation of TAI; after all, it's a notion that has been invoked in many characterisations of AGI and HLMI.

However, traditional LLMs becoming more 'generally intelligent' in this sense without also possessing a high degree of generality will require *large amounts of data*. (This is because, if traditional LLMs cannot apply learnings from their training distributions to new problems, their capabilities depend on what they have 'memorised' from their training distributions, and improving their capabilities will require making them memorise more.) As such, adopting this line of response makes us more vulnerable to the 'data as a bottleneck' arguments highlighted earlier.

"Even general intelligence isn't necessary for transformative-level capabilities." Suppose we take the argument of the previous bullet point to imply that traditional LLMs will not become fully general unless we can solve some data challenges. But even if AI systems never become fully general, couldn't they still be transformative?

The sorts of things that traditional LLMs are currently able to do, whether or not we see them as evidence for increased general intelligence (or for increased generality), seem important in their own right. For example, improvements to the skills that these systems already have, even if limited to relatively narrow domains, could still lead to the automation of critical economic tasks. Moreover, the narrow capabilities of individual systems could become especially powerful if composed together in the form of a network.

"Sceptics have been proven wrong before!" In his essay on *Situational Awareness*, Leopold Aschenbrenner (citing examples from Yann LeCun, Gary Marcus, and Bryan Caplan) writes: "Over and over again, year after year,

skeptics have claimed 'deep learning won't be able to do X' and have been quickly proven wrong. *If there's one lesson we've learned from the past decade of AI, it's that you should never bet against deep learning.*"

Further reading:

- *François Chollet on Dwarkesh Patel podcast (plus this summary)*
- *On the measure of intelligence - François Chollet*
- *The Debate Over Understanding in AI's Large Language Models - Melanie Mitchell*
- *ARC prize website*
- *LLM generality is a timeline crux - EggSyntax*
- *OpenAI o1 Results on ARC-AGI-PUB - Mike Knoop*
- *OpenAI o3 Breakthrough High Score on ARC-AGI-PUB - François Chollet*
- *OpenAI's o3 isn't AGI yet but it just did something no other AI has done - Tiernan Ray*
- *Five breakthroughs that make OpenAI's o3 a turning point for AI — and one big challenge - Matt Marshall*
- *AI Capabilities Can Be Significantly Improved Without Expensive Retraining - Epoch*

The environment as a bottleneck

This might be viewed as a second sense in which data could be a bottleneck for AI capabilities progress, but it involves very distinct considerations (and features less commonly in the literature) and is therefore worth highlighting separately. Here, I'm no longer concerned with the *amount of high quality data* available to train an AI, but with something more complex: the specific *design* of the training environment (i.e. the overall ecosystem in which the AI system is trained, which includes its dataset as well as the training algorithms deployed).

A sceptic of short timelines via compute scaling might argue that no matter how much compute we use to train an AI, transformative capabilities are unlikely to emerge unless the training environment actually incentivises their development. So the question is: how difficult will it be to design a training environment that is conducive to the development of TAI? Could this be a significant bottleneck?

Richard Ngo has introduced the hard paths hypothesis to capture the view that *it's rare for environments to straightforwardly incentivise the development of general intelligence*.

If you believe this hypothesis (and view something like 'general intelligence'[41] to be necessary for TAI), then you will probably believe that short TAI timelines are implausible – unless you take some specific inside view which suggests that AI developers have already hit upon one of those rare training environments in which general intelligence is incentivised, or will soon do so.

[41] In the sense that's most relevant here, 'general intelligence' refers not specifically to the 'ability to generalise' to novel problems, but to the (strongly related) ability to perform well on a broad range of tasks (as opposed to just a narrow domain). As mentioned earlier, this latter ability is common to many conceptions of TAI.

It's not obvious whether the hard paths hypothesis is true or false. Of course, we know it's *possible* for an environment to incentivise the development of general intelligence; the evolutionary environment in which human intelligence emerged is an example of this. But this does not provide us with any indication of how *commonly* such environments occur in reality, or how *easy* they are to design.

Below, I note some arguments in favour of the hard paths hypothesis being true:

- In the history of AI development so far, it's been relatively easy to make improvements within narrow domains, but we've struggled with getting general intelligence. Abilities that we previously thought were marks of high levels of general intelligence – such as the ability to perform well at games like Chess and Go – proved not to be. Indeed, achieving expert-level performance at such games did not require AI systems to develop any skills *outside* of those games; it seems their training environments only incentivised improvements within very narrow domains.
- Perhaps there was some really difficult step in the evolution of human intelligence that environments very rarely enable. This might be something akin to Robin Hanson's 'Great Filter'.
- Since the development of human intelligence was so strongly tied to the interaction of humans with a physical world, perhaps interacting with some *detailed simulation of* the world will be necessary for the development of general intelligence in AIs. This could be quite a difficult (and resource intensive) training environment to provide.

There are also reasons to think this argument is false. For example, note the smoothness of empirical scaling laws: the capabilities of AI systems have been consistently and predictably improving with the scale of their training runs. This suggests that it is not common for training runs to get stuck in local minima indefinitely.

For more on this particular debate, see the further reading listed below.

Further reading:

- *Environments as a bottleneck in AGI development - Richard Ngo*
- *Section 3.2.1.1 of Modelling Transformative AI Risks - Clarke et al.*

Embodiment as a bottleneck

Some people believe that achieving TAI will require AI systems to be, in some sense, embodied.

Proponents of this view observe that the development of human intelligence (and intelligence in other animals) occurs through a series of rich interactions between the being and the world, mediated by a body. With this in mind, they argue that embodiment is, in fact, a central component of intelligence, or even a prerequisite for its emergence. Such views have gained traction over the last

few decades through the embodied cognition movement, and have been influenced by the earlier work of psychologists from the ecological and connectionist traditions, as well as phenomenologists such as Merleau-Ponty.

Arguments of this kind about human intelligence can be extended to the case of AI development as follows: if advanced forms of artificial intelligence are considered to be anything like human intelligence[42], then perhaps developing TAI will *also* require enabling AI systems to interact with a world, mediated by a body, in a similar way to humans.

The type of AI 'embodiment' needed for TAI to emerge could be construed in a few different ways:

Virtual embodiment. Maybe embodiment, in the relevant sense for AI development, is something that could be captured in the form of a dataset and training environment. In the previous subsection, I indicated the possibility of creating a simulation of the physical world to train an AI on; this may be viewed as a virtual equivalent to being embodied that would provide the rich interactions necessary for the development of higher levels of intelligence.

I've already noted that creating a detailed world simulation with rich scope for interaction might be difficult and resource-intensive. In relation to the hard paths hypothesis, Richard Ngo posited in 2020 that "implementing flexible interactions in simulations will be very difficult – even state-of-the-art video games are far from supporting this". Although there have certainly been recent efforts in the direction of creating virtual worlds, it does seem that achieving this within the next ten years is an ambitious ask, unless a fairly low-fidelity simulation would suffice for the purposes of training TAI.

Physical embodiment. Perhaps 'embodiment' in the sense that will be necessary for TAI is more or less the same as in the biological case: that is, AI systems possessing physical bodies. One could argue that this is the only feasible way for AI systems to enjoy the necessary kinds of interaction with the world, or to experience this world to a high enough degree of fidelity, for developing TAI-level capabilities. This might involve equipping AI systems with robotic limbs and sensors, through which they can navigate their physical environment and learn by receiving feedback from it.

Notably, and unlike in the case of virtual embodiment, this view emphasises an input to AI progress that falls decidedly *outside of the Triad* of compute, data, and algorithms. The key implication of this view, in the context of this chapter, is: *no amount of compute* (even if all the data and algorithm bottlenecks discussed above are overcome) will get us to TAI unless accompanied by improvements in robotics, and the integration of robotics with AI.

Could this be achieved within the next ten years? Given that investment and effort within the field of AI development is (at present) predominantly directed towards 'disembodied' systems, it seems unlikely. However, if the field were faced with increasing motivation to explore robotics as a route for advancing AI capabilities (e.g., if progress in the capabilities of disembodied AIs began to stall, prompting new solutions to be pursued), there's no *obvious* reason that

[42] That is, if artificial and human intelligence are expected to be similar in a sense that has a bearing on the way it is developed or instantiated. This position could be pursued from a few angles. (1) We could note that TAI-level capabilities might consist in the ability to perform certain 'human tasks', as gestured to in popular definitions of AGI, HLMI, superintelligence, and so on. Perhaps the ability to perform the particular tasks that humans do will require developing intelligence of a similar kind, or through a similar route. (2) More generally, we might believe that there are certain fundamental features of what we call 'intelligence' which are necessarily shared by *any* 'intelligent' (or highly capable) being, regardless of whether such beings are biological or non-biological. This view gestures towards some universal notion of *what it means to be intelligent*, rather than offering examples of similarities between the specific cases of human intelligence and TAI.

suitable mechanical bodies couldn't be developed and integrated with AI systems fairly quickly. This would somewhat reduce the likelihood of a short timeline, but certainly not rule it out.

In any case, views of AI progress that emphasise physical or virtual embodiment have been waning in popularity over the past few years. This is largely due to the recent successes of disembodied AIs, which are taken to illustrate that embodiment (in either sense) is not necessary for AI systems to become highly capable at a range of important tasks. In light of this, I won't discuss this subject further here.

Further reading:

- *Why Heideggerian AI failed, and how fixing it would require making it more Heideggerian - Hubert Dreyfus*
- *'Fallacy 4' in Why AI is harder than we think - Melanie Mitchell*
- *SEP entry on Embodied Cognition*
- *Wikipedia entry on Embodied Cognition*
- *Aligning Cyber Space with Physical World: A Comprehensive Survey on Embodied AI - Liu et al.*

Category 2 sceptical arguments: The TAI compute threshold is out of reach within the next ten years

Suppose you accept the compute-centric scaling hypothesis: that is, you believe scaling up compute for current systems is the main thing that's needed to achieve TAI, and that none of the potential bottlenecks identified above will be a major obstacle to AI progress *if developers have access to enough training compute*.

You might still believe that TAI will not emerge by 2035 simply because *developers won't have enough training compute by that point*.

A sceptical argument of this kind could be made from two complementary angles:

(i) Compute growth just won't be that fast.

(ii) The compute threshold for TAI is really high.

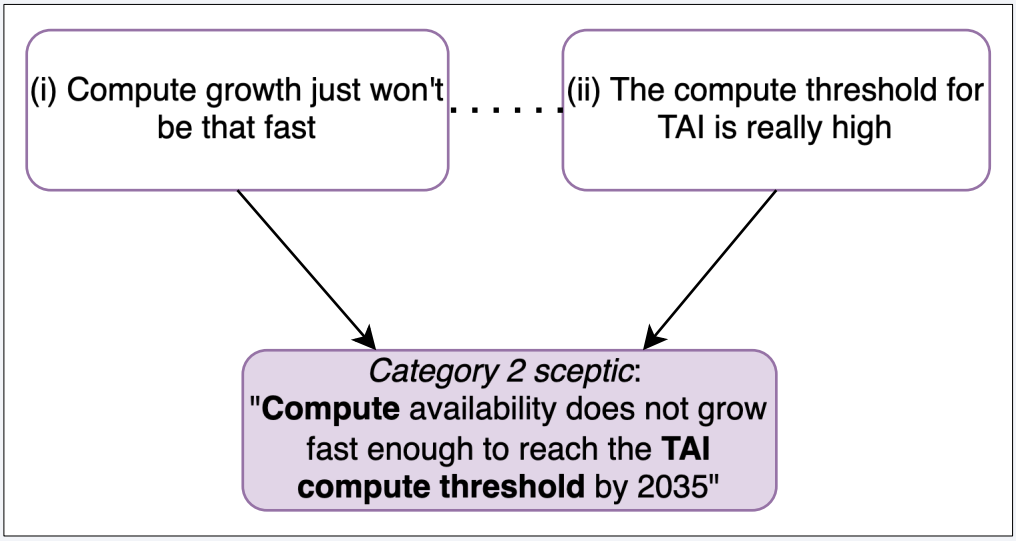


Figure 1.9: This represents the structure of the arguments in this subsection. As before, the dotted line indicates that claims (i) and (ii) are best understood in relation to one another.

Compute growth just won't be that fast

Growth in the training compute available to developers could be constrained or slowed in a number of ways.

Physical constraints on chips

Perhaps Moore's Law has already ended or will soon end, as it comes up against physical constraints.

Of course, as I noted previously, improvements in other areas (such as the number of chips used, chip design, and so on) could still continue to push compute growth forward in this case. Nonetheless, the death of Moore's Law would remove one important driver of compute growth.

Power as a bottleneck[43]

Large-scale training runs come with huge power requirements. For example, Aschenbrenner estimates that scaling today's largest training clusters up by 3 or 4 OOMs of compute – which might be what's necessary for a single TAI-scale training run – will require 10s or 100s of gigawatts of power.[44] For reference: the US currently consumes less than 500 GW of electricity each year, and has about 1,300 GW of installed generation capacity. 10GW would be akin to the yearly electricity consumption of an entire state, and 100GW to that of multiple states. With this in mind, Aschenbrenner notes that power requirements for training are "probably the single biggest constraint [to compute scaling] on the supply-side".

A recent Epoch report titled Can AI scaling continue through 2030? attempts to quantify this constraint to compute scaling. It finds that, out of a list of four potential constraints (power, manufacturing, data, and latency), power requirements are the most likely to present a bottleneck to scaling up training compute over the next five years.

Specifically: Epoch estimates that if compute growth continued to follow

[43] As of January 2025 (after this report was finished), some of the arguments below are now outdated in light of a series of executive orders by President Trump. In a few subsequent footnotes, I include details of the key changes to be aware of. Overall, power seems like less of a barrier for developing TAI now than it did at the original time of writing, but is still worth factoring into the debate.

[44] That's just for training alone; running inference on a trained model will push power requirements up even further!

scaling trends, unconstrained by power requirements (and other potential bottlenecks), training runs of 2*10^29 FLOP would be achievable by 2030. The authors predict with around 50% confidence that the largest possible training run by 2030 will fail to meet this compute projection if constrained solely by power requirements, as illustrated in the graph below.

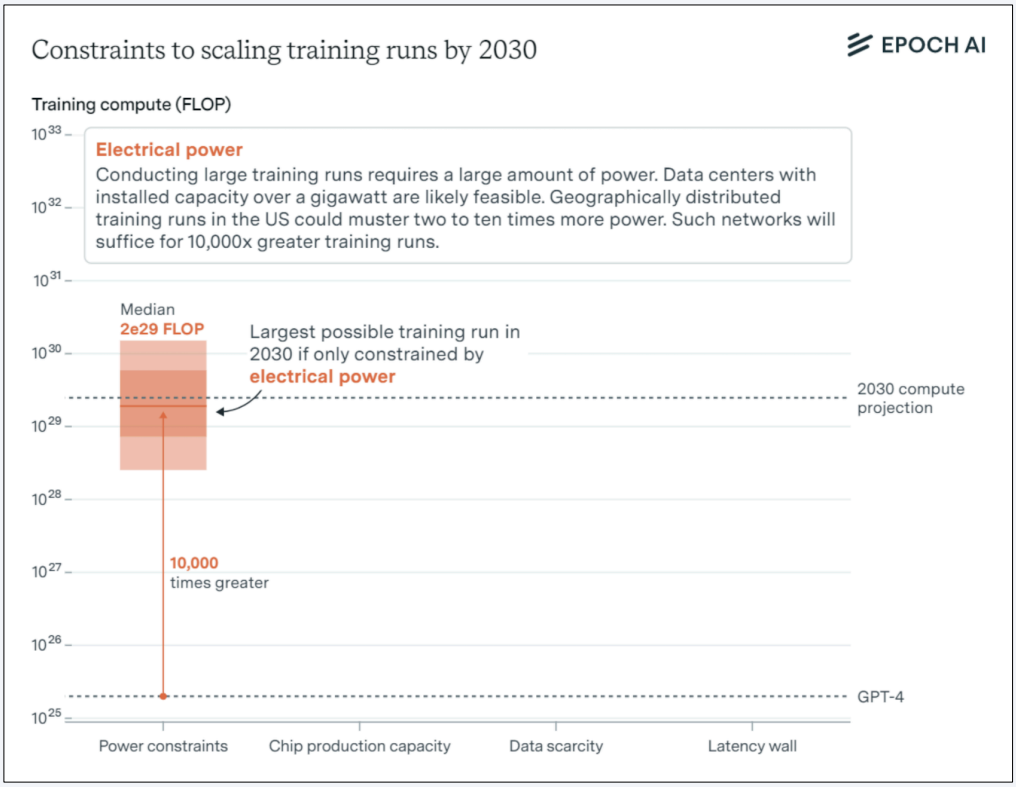


Figure 1.10: Compute scaling vs power constraints. This box plot graph quantifies the potential for power requirements to create a bottleneck to scaling training runs over the next five years. *Source: Epoch.*

Looking at the graph, the potential size of this bottleneck is fairly small: the error bars indicate that if we fail to meet the compute projection for 2030, it will probably be by less than one order of magnitude of FLOP. Still, even this is enough to cast real doubt over the plausibility of powering a TAI-scale training run by 2035.

Despite the challenge to TAI development presented by power requirements, Aschenbrenner still believes a TAI-scale compute cluster could be powered in the US within the next few years, using natural gas. Specifically, he suggests that the US could secure enough power for:

- A ~10GW cluster by ramping up, or redistributing, the production of existing natural gas facilities (since 10GW is only a small percentage of current natural gas capacity in the US); and
- A ~100GW cluster by drilling around 1,200 new wells and building new generators and turbines.

There are serious barriers to this approach.

Firstly, the creation of significant additional power infrastructure for the

purpose of supporting an AI project, as in the 100GW cluster scenario, would be a lengthy process. Like all major infrastructure projects, before construction can even begin, there would be years of planning, environmental assessments, permitting, commissioning, procurement, contractual and financial agreements, and so on. These processes are especially thorough in the US, and are often roadblocked by legal disputes, objections from local communities, and red tape. It seems unlikely that a project of this scale would be completed within the next ten years.

At the current time of writing, it also seems that climate commitments (from the US government, and from AI labs themselves) and broader pressures to decarbonise might present difficulties for any solution emphasising natural gas, even in the 10GW cluster scenario.[45] (However, it is worth noting that the US is still pursuing a significant build-out of natural gas capacity in the short term, so the idea of ramping this up just enough to support an additional 10 GW of power production is not an unreasonable stretch of the imagination. In fact, existing plans to increase gas capacity have been partly motivated by a surge in demand from data centres.)[46]

Aschenbrenner acknowledges these obstacles, but argues that they're "entirely self-made", and could basically be dissolved if the rapid creation of TAI is deemed important enough for US national security. However, *as of the time of writing* (late 2024), there's no clear evidence that decision-makers are on the brink of going to any extreme lengths to expedite the process of developing TAI, at the expense of the usual protocols and environmental considerations that projects of this scale of disruption are ordinarily subject to. Perhaps there will soon be a shift in mood at the level of government, especially with a new US administration coming into power in 2025; I do not offer any speculation on this point here.[47]

Natural gas certainly isn't the only option (or even the default option) for powering large-scale training runs in the US. There's growing interest in the use of nuclear energy for this purpose, with Google, Microsoft, and Amazon all having taken recent steps to secure nuclear power sources for their data centres. However, many of the same obstacles highlighted above for scaling up natural gas power production also apply in the nuclear case: it would still be a lengthy process (probably even lengthier!) to commission new facilities, and nuclear projects come with their own serious set of environmental concerns, safety considerations, and public perception issues. Even with the potential to bring decommissioned or offline nuclear reactors back online (rather than setting up a new facility from scratch), as Microsoft is currently exploring with its Three Mile Island plans, it doesn't seem likely that the nuclear route for powering TAI training runs would be any *faster* than Aschenbrenner's proposed natural gas route.[48] It's hard to make a stronger judgement than this on the comparative timelines, since there are so many situation-specific variables which determine the timeline for implementing any given power project. But in any case, nuclear power seems to be the direction that leading tech companies are currently pursuing.[49]

[45] Note that after this report was written, the Trump administration announced that the US would be leaving the Paris Agreement, and declared its intentions to boost oil and gas power production. In Trump's inaugural words: the US will "drill, baby, drill". (20th January 2025.) It is still true that there are existing *pressures* to decarbonise, but it's not clear that these pressures will actually present much of a barrier for developing TAI under the new administration.

[46] Indeed, there's growing concern that power demand from AI will seriously undermine efforts to transition to cleaner energy (this has been discussed e.g. in The Wall Street Journal).

[47] The contents of this paragraph have been largely outdated by Trump's series of executive orders, announced in his inaugural address in January 2025. As highlighted in a previous footnote, the new administration has announced that the US will leave the Paris Agreement and ramp up on oil and gas. Trump has also signed an executive order on AI revoking policies from the previous administration which he believes "act as barriers to American AI innovation." The new administration's plans include a $500 billion project with OpenAI (the 'Stargate' project) to construct the data infrastructure required for developing advanced AI. It now seems fair to say that decision-makers are (in my earlier words) "on the brink of going to ... extreme lengths to expedite the process of developing TAI", based on a "shift in mood at the level of government".

Of course, the US is not the only country that could potentially train a TAI system in the next decade. And in some other parts of the world, bringing online the necessary power infrastructure for a TAI training cluster might be considerably quicker and easier to achieve.

Investment as a bottleneck

All drivers of compute growth require investment. Indeed, the compute growth we have observed over the last eight years has relied on exponential increases in spending on training runs, to the tune of 2.4x each year since 2016 (according to Epoch). For current compute trends to continue, there has to be enough money available to developers, as well as the economic motivation to funnel that money towards scaling systems up. But what if there's not enough money or economic motivation to sustain fast scaling?

Scaling to TAI would likely require spending vast amounts of money on training runs. For example, Aschenbrenner estimates that training AGI will cost 100s of billions of dollars, and acknowledges the possibility that it might even cost trillions. The investment projections used in quantitative forecast models for training TAI, such as Epoch's Direct Approach, are in a similar ballpark range.

One might doubt that projects this capital-intensive will be a possibility for developers any time soon. Or, even if it's possible, perhaps the motivation to direct massive investment towards increasing scale just won't be there: after all, there may be a better business case to pour resources into making models with *current levels of compute* smaller and cheaper. In this vein, Narayanan and Kapoor argue that:

> "capability is no longer the barrier to adoption… there are many applications that are possible to build with current LLM capabilities but aren't being built or adopted due to cost, among other reasons."

However, these objections don't seem to reflect the current zeitgeist around AI development. Although there may be some efforts towards making existing systems cheaper, it appears there is significant interest still in the direction of huge investment, (likely) compute-guzzling systems – and there are no obvious signs of this stopping in the near future. Indeed, in Chapter IIIa of *Situational Awareness*, Aschenbrenner notes that:

> "Zuck bought 350k H100s. Amazon bought a 1GW datacenter campus next to a nuclear power plant. Rumors suggest a 1GW, 1.4M H100-equivalent cluster [likely requiring $10s of billions]… is being built in Kuwait. Media report that Microsoft and OpenAI are rumored to be working on a $100B cluster, slated for 2028 (a cost comparable to the International Space Station!). And as each generation of models shocks the world, further acceleration may yet be in store."

It seems that training runs costing in the 100s of billions of dollars are a genuine near-term prospect.[50] However, Aschenbrenner does still acknowledge that achieving the first $1 trillion training run will require "a truly

[48] For example, Microsoft's proposed Three Mile Island project would take years to get up and running, given the need for regulator approvals, detailed environmental and safety reviews, and (likely) plant upgrades/refurbishment before the nuclear reactor can be brought back online. Further delays or challenges could come from public opposition, given that the plant in question was the site of the most serious nuclear meltdown in US history.

[49] The decisions by AI labs to opt for nuclear power rather than, say, natural gas are likely due to a combination of economic factors and a need to keep in line with their climate commitments. Indeed, Microsoft called its plans to reopen Three Mile Island a "milestone" in its efforts to "help decarbonize the grid".

Notably, there's no evidence that the *speed* of implementing projects has played any part in motivating these decisions.

[50] Note that in January 2025, after this report was written, Trump announced a $500 billion project with OpenAI ('Stargate') to construct the infrastructure required to support the development of advanced AI, as well as a separate $20 billion into datacentres. This shows a clear willingness on behalf of the US government to invest into large-scale advanced AI, and should increase our degree of belief in the claim that challenges on the investment side will soon be overcome.

extraordinary effort", possibly even involving the work of a national consortium. Although there are historical precedents for $1 trillion spending into other infrastructure (e.g. internet, green energy), this would be a big step for the AI field, and it's certainly not *obvious* that we're on track for it to happen soon. So, if the total investment that is required for training TAI is actually to the order of trillions of dollars rather than 100 billions, there is still some serious uncertainty over whether this will even be a financial possibility by 2035.

Manufacturing and supply chain

With sufficient investment, AI labs will be able to buy more and more computational power. But as demand for compute grows, the semiconductor supply chain will eventually reach a point where it just *cannot move fast enough* (or *does not have enough manufacturing capacity*), to accommodate the provision of these huge amounts of compute. This is because, as Ajeya Cotra highlighted in her recent dialogue with Ege Erdil and Daniel Kokotajlo:

> "you need to build more datacenters which come with new engineering challenges, more chip-printing facilities, more fabs, more fab equipment manufacturing plans, etc."

These supply chains can only operate so fast. This is likely to put hard limits on compute growth rates. (And all it takes to bottleneck compute growth is for *one component* of the supply chain to be unable to keep up with demand; after all, the chain can only move as fast as its slowest component.)

So, the question of *when the supply chain reaches this point* is another crux for timeline projections. If you think it will happen before TAI is developed, you're likely to expect longer timelines to TAI. Indeed, in the aforementioned dialogue, both Cotra and Kokotajlo endorse the (compute-centric) scaling hypothesis, but nonetheless make quite different timeline projections; it seems that their varying opinions on this particular issue partially account for that difference.

Market dynamics

Even if there's enough compute available to train a TAI system *in theory*, market pressures may prevent such a system from *actually being trained* in the near future. For example, it's possible that competition between AI developers will increase such that no single actor can get hold of a high enough proportion of the total available compute to develop TAI within the next ten years.

Governance and regulatory landscape

In the discussion so far, I've largely assumed business-as-usual on the side of AI governance. However, certain interventions (such as the introduction of direct restrictions on compute access) would very likely slow down compute growth. Rather than elaborating on this point here, those interventions are treated as potential 'levers' for AI governance in other research pieces by the Convergence team.

The compute threshold for TAI is really high

We don't know how much compute would be sufficient to train a TAI-level system. Attempts to estimate this have typically taken the form of a wide probability distribution. For example, in her Biological Anchors report, Ajeya Cotra aggregates six probability distributions for compute requirements corresponding to six hypotheses about intelligence, and notes that:

> "The resulting distribution is very wide. Conditional on one of the hypotheses being true, the distribution places non-trivial probability mass on a range of **26 orders of magnitude**, from **1e24 FLOP to 1e50 FLOP**". (Bolding is mine, for emphasis.)

It might be that the compute *actually* required to train a transformative system is on the higher end of this spectrum. This would make it highly implausible for compute scaling alone to produce TAI by 2035: the largest training run to date has been to the order of 10^25 FLOP[51], and there's no clear route for obtaining an additional 25 OOMs within the next decade[52]. (And, even if there were, the residual heat from powering 10^50 FLOP might just boil the oceans!)

One might try to at least place a rough upper bound on compute requirements for TAI. For example, since we know that evolution gave rise to human intelligence, we could argue that the total amount of computation performed in evolution would also be sufficient for training human-level AI systems (Cotra's median estimate of this quantity is 10^41 FLOP). However, in *Superintelligence*, Nick Bostrom points out that it's difficult to derive any meaningful upper bound for compute thresholds based on such evolutionary arguments. The evolution of human intelligence might not have simply required the total computational power of evolutionary processes, but also a great deal of *lucky coincidence*. Perhaps only in one of 10^30 worlds would 10^41 FLOP of computation even give rise to human-level intelligence; maybe in other situations, even more FLOP than this would be required. So: even if we managed to replicate the total computation performed in human evolution to train an AI system, there's no guarantee that it would end up with anything close to human-level intelligence.[53]

Taking stock

Compute scaling is evidently a powerful mechanism for AI capabilities progress, and is likely to continue – at least in the near term – to play a central role in the development of more advanced AI systems. However, it's not clear exactly *how far* compute scaling will take the current AI paradigm over the next decade.

The counterarguments of this chapter illustrate some reasons to doubt that compute scaling with LLMs will get us all the way to TAI by 2035; it seems likely that, at some point, challenges will emerge and have a slowing or constraining effect on existing trends. Firstly, capabilities progress could get bottlenecked on something other than compute, and gains from compute

[51] Specifically, Epoch has reported that "The total training compute for the final training run of Gemini Ultra, likely the most compute-intensive model to date, is estimated at 5e25 FLOP."

[52] (At least, there's no clear route that doesn't involve some form of recursive improvement dynamics, which are discussed in the next chapter.)

For an illustration of how far out of reach an additional 25 OOMs of training compute seems to be in the next decade, see Epoch's report Projecting Compute Trends in Machine Learning. Here, 10^50 (physical) FLOP is not even within two standard deviations of the median compute projection for 2080, let alone for 2035. Projecting effective FLOP (i.e. taking improvements in algorithmic efficiency into consideration) rather than physical FLOP would likely only yield a *slightly* more favourable result.

[53] This could be posed as a variant of, or corollary to, the 'hard paths hypothesis' I discussed earlier. Recall: the sceptic who endorses the hard paths hypothesis believes that environments that are conducive to the development of certain levels of intelligence are rare. To extend her argument to this context, she might suggest that an environment 'being *more conducive* to developing TAI' would mean something like 'making TAI achievable with *less training compute*'.

However, I think this point from Bostrom is better posed in the present context than in relation to the idea of 'hard paths'. According to my earlier framing, the hard paths argument identifies a specific bottleneck for AI capabilities improvements – the design of the training environment – which AI developers will try to overcome, but which might delay the arrival of TAI. By contrast, Bostrom gestures broadly towards a sheer matter of 'lucky

coincidence' here, rather than some difficulty posed by a specific, controllable input to AI development, like the design of a training environment. In other words, he isn't highlighting a specific bottleneck to AI progress. And importantly, since this element of 'lucky coincidence' is unquantifiable, it points to significant uncertainty over *any* proposed upper bound for compute requirements, and therefore over any timeline projections based on such estimates.

54 As noted earlier in this chapter, other forms of compute scaling may still drive rapid improvements in performance even if the specific strategy of increasing training compute diminishes in effectiveness.

scaling could diminish or stall as a result (in fact, there's some recent evidence that suggests this has already started happening *on the side of training compute*[54]). Secondly, developers might struggle to keep scaling up their frontier AI models due to constraints around compute availability or use. If either of these situations occur *before* TAI has arrived, they might seriously impact the likelihood of a short TAI timeline.

However, given the current pace of capabilities progress, it's possible that TAI will already have emerged before these issues seriously affect the field. And even if compute trends do break down in the near future, there are other mechanisms that could kick in through which capabilities progress could still be very fast. In particular, some form of *recursive improvement* could effectively break bottlenecks or lift capabilities progress out of a plateau, as we'll see in the next chapter.

Chapter 2: Recursive improvement

NOTE

Although 'recursive improvement' and related terms such as 'intelligence explosion' are commonly used to describe the dynamics of takeoff from AGI to ASI, this chapter will continue to focus on the development of the first transformative AI systems, whatever form they take. That is, the recursive improvement stories characterised below cover progress to TAI from what is assumed to be a pre-transformative (and therefore pre-AGI) capabilities threshold.

What is recursive improvement, and how could it produce a short TAI timeline?

By 'recursive improvement' in AI development, I broadly mean any iterative process characterised by feedback loops through which there are *repeated improvements to the ability to improve AI*.

Feedback loops and growth modes

There are many examples of systems which have exhibited positive feedback loops. These include: population; world GDP; some financial markets; the process of biological evolution; progress in specific technological fields such as DNA sequencing, computing, and the internet; and scientific progress as a whole. These loops are often associated with exponential growth modes[55] – and sometimes even hyperbolic growth, as in the case of population and world GDP. However, not *all* positive feedback loops bring about very rapid improvements in their domain; for example, improvements may only grow as a slow exponential or, due to bottlenecks, flatten to a sub-exponential trend.

Some experts anticipate short TAI timelines because they expect that, at some point in the decade ahead, positive feedback loops in the field of AI development will bring about a period of *fast* recursive improvement in AI capabilities.

There are varying beliefs about what the trajectory of AI development would actually look like in this scenario, but it is often imagined that the first TAI (and even more capable systems) would emerge through some "sustained period of accelerating growth"[56] in AI capabilities. This is sometimes described

[55] Indeed, in *The Singularity Is Near*, Kurzweil argues that biological and technological systems typically exhibit exponential growth modes, as a result of positive feedback loops. This has been dubbed Kurzweil's 'law of accelerating returns'.

[56] This quotation has been taken from David Thorstad's summary of the singularity hypothesis in *Against the singularity hypothesis*.

In this report, I often use the word 'sustained' to differentiate *a period* of accelerating growth from *a single step change* in growth rate. Creating TAI within the next ten years would not necessarily require sustained acceleration in capabilities growth; as will become clear later on in this chapter, a step change in growth rate may well be sufficient (or more than sufficient) to get us there by 2035. However, many of those who expect short timelines through recursive improvement believe, and build into their stories, that the acceleration of AI capabilities progress will be sustained – that is, we'll see repeated increases to the rate of progress over some period of time.

as an *intelligence explosion* or *singularity*, though terms are used variably in the literature.[57]

Chapter roadmap

In 'Categories of recursive improvement, and the route to TAI', I will outline different types of recursive improvement that AI R&D could exhibit (or, indeed, already exhibits), and explain how each of these mechanisms could speed up AI progress. This will include some discussion of the possible trajectories of recursive improvement on the pathway to TAI.

Next, in 'Overcoming scaling bottlenecks via recursive improvement', I will explain how positive feedback loops of the outlined kinds could help the AI field overcome some of the bottlenecks to scaling identified in the previous chapter.

When I turn to explore 'Why recursive improvement might not produce a short TAI timeline', I'll be especially concerned with evaluating stories of what I call **'direct' recursive improvement (DRI)**. Here, the primary question of interest will not be *whether AI R&D will exhibit direct recursive improvement or fail to do so*. Instead, I'll be concerned with *whether such recursive improvement dynamics would produce a short TAI timeline* if realised.

The answer to this question is determined by four variables:

(1) How **fast** the resulting growth in capabilities would be;

(2) How **long** this growth period could be sustained for;

(3) **When** this growth period **begins**; and

(4) **How far** away TAI is at the point at which it begins.

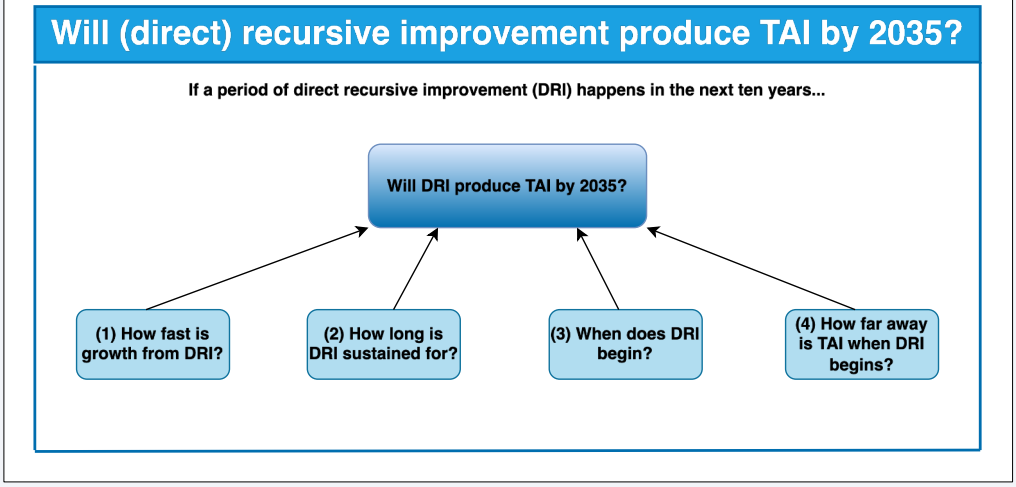


Figure 2.1: Will direct recursive improvement produce TAI by 2035? The above graphic illustrates the four variables determining the answer to this question (in worlds where DRI dynamics do emerge).

I'll pay particular attention to variables (1) and (2) on this list. Specifically, I'll lay out and analyse *three possible arguments* through which a sceptic might resist the idea that (1) direct recursive improvement will result in extremely rapid capabilities growth, or that (2) such rapid growth could be sustained for

[57] The term 'singularity', for example, is also sometimes used to describe a particular event horizon (rather than a period of time): namely, it can refer to the specific point in time at which technological/AI progress becomes uncontrollable and irreversible.

long enough to bring about TAI.

In 'Who wins the tug of war?', I'll evaluate the implications of these three counterarguments on the plausibility of short TAI timelines. Here, I'll consider what trajectory of capabilities improvements would actually be necessary for a short timeline to be realised, and illustrate that even if many of the specific challenges identified by the sceptic are taken into consideration, recursive improvement dynamics could *still* be argued to bring about TAI within the next ten years.

I will end with some discussion of variables (3) and (4) on the list above, though this will be less detailed than the corresponding discussions for variables (1) and (2).[58]

A final note: A lot of new terminology is introduced in this chapter, especially in reference to 'direct' recursive improvement. A **glossary of key terms** relating to direct recursive improvement can be found in Figure 2.2.

Two categories of recursive improvement, and the route to TAI

There are many possible feedback loops through which AI capabilities progress could exhibit recursive improvement. Some of these might be described as 'direct', in the sense that AI systems themselves play a central role in improving the ability to improve AI. Those which are primarily mediated by influences *external* to the AIs themselves might be described as 'indirect'. This is not a rigorous or clean division, but will be instructive for this chapter.

Put more simply, the distinction I'm gesturing towards might be seen as follows:

- In **direct recursive improvement**, AI capabilities improvements are driven by AI systems.
- In **indirect recursive improvement**, AI capabilities are improved via broader societal feedback loops.

Direct recursive improvement

When people talk about 'recursive improvement' in AI development, they are typically referring to improvements mediated by 'direct' feedback loops of the kind described above. Likewise, terms like 'intelligence explosion' are usually associated with *direct recursive improvement scenarios* rather than indirect ones.

There have been many influential models, characterisations, and arguments for direct recursive improvement scenarios in AI development. These include works from I.J. Good, David Chalmers, Nick Bostrom, Eliezer Yudkowsky, Eric Drexler, Tom Davidson, and Leopold Aschenbrenner.

Broadly, they all share the following basic structure:

[58] Variables (3) and (4) will be of some interest in the scenario generation process of Chapter 3, as we'll see.

- AI systems (or a single system) take on a central role in improving AI capabilities.
- With each improvement to AI capabilities that is made, these systems *get better* at driving future AI capabilities improvements.
- A 'direct' feedback loop thus emerges.

As the breadth of different characterisations in the literature indicates, there is a rich variety of ways that this basic story could unfold, each emphasising different types of direct feedback loops that could emerge. Some stories involve AI systems improving their *own capabilities*, while others involve the AIs developing *successor systems*. Some highlight improvements spanning *across the field* as a whole, while others highlight improvements within a *specific domain* or on a *specific input* to progress. Some maintain that *significant human involvement* is necessary for supporting these improvements, while others do not.

Moreover, as time goes on, developers are discovering promising *new* avenues for involving AI systems in making AI capabilities improvements, and thereby introducing direct feedback loops. For example, the idea of AIs which iteratively improve their own performance via reinforcement learning is gaining legitimacy as an option for near-term AI progress, given recent successes in the use of such methods (for example, in OpenAI's o3 model).

Though all of these stories have great structural similarities to one another, each different type of direct feedback loop we could imagine here has its own likelihood of emerging within the next few years, as well as its own consequences for the subsequent trajectory of AI capabilities progress.

Focusing on stories of AI R&D automation

For the purpose of this report, I want to draw our focus towards versions of the basic story above in which a fleet of AI systems automates some significant proportion of work in the AI R&D field. With each contribution to AI R&D made by these systems, the next generation of systems is more capable than the last, and therefore better placed to make even further contributions to the field; a direct feedback loop thus emerges. (Under the next subheading, I'll walk through a story of this kind in more detail.)

Of the variety of possible feedback loops, these 'AI R&D type' direct feedback loops seem to have the most potential to strongly affect capabilities progress *before the arrival of TAI*, which makes them especially relevant in the context of short timelines to TAI.[59]

Stories featuring AI R&D type feedback loops have dominated in the literature on direct recursive improvement over the past few years. This is likely due to the release of GPT-2 in 2019 (which was taken by many to show early signs of the potential for generative models to contribute to AI R&D), alongside Drexler's publication of *Reframing Superintelligence* in the same year (which advanced one of the first major arguments for recursive improvement via AI R&D automation).

[59] More precisely: the claim I'm making here is that, of those direct feedback loops in the AI field which seem likely to emerge before TAI has already been developed, these 'AI R&D type' loops have the strongest chance of yielding very fast improvements to capabilities. However, I don't do an explicit comparison here of the relative likelihoods and speeds of improvement associated with different types of feedback loops.

From now on, *when I talk about 'direct recursive improvement' in this report, I'm referring to* ***recursive improvement as defined by 'AI R&D type' direct feedback loops***.

Deliberately, this characterisation still admits multiple interpretations. I won't always distinguish between these interpretations – but when I do, I'll frame things around the idea of *automated workers* in AI R&D, which aligns particularly well with ideas from Leopold Aschenbrenner and (perhaps to a lesser extent) Tom Davidson.

THE *AUTOMATED WORKER* FRAMING

Imagine that we have a fleet of AI systems, each of which can perform *all or most* of the work done by a researcher/engineer in the field of AI R&D. These systems are deployed as 'drop-in replacements' for workers in the field. That is: they complete *more or less* the full range of tasks typically performed by these workers to the same standard or higher, doing so end-to-end (i.e. from idea generation to implementation) and without *significant* reliance on human supervision or prompting.

Note that qualifiers like '*more or less*' are doing some work here. AI systems don't need to perform exactly 100% of the tasks typically performed by researchers/engineers in the AI R&D field – or do so with exactly *zero* human support – to be reasonably said to have 'replaced' those human workers.

My rough characterisation of 'automated workers' says nothing about what the salient threshold actually is here. (And in fact, what I'm gesturing towards here might be best envisioned as a spectrum, anyway.) In what follows, I'll continue to be vague on this subject – for example, by describing these workers as being able to perform 'all or most' tasks in AI R&D or 'a significant proportion' of work in the field, without 'substantial' human involvement.[60]

I also assume for the sake of this chapter that the relevant capabilities threshold here – whatever it is exactly – is *pre-transformative*. (Otherwise, the present argument will be irrelevant in the context of timelines *to the arrival of* transformative AI.)

I think this is a plausible assumption to go forward with.[61] The above framing of automated AI R&D workers does come apart, at least conceptually, from most popular characterisations of TAI. In particular, these workers do not need to excel at *as wide a range* of economically or cognitively relevant tasks as, say, an AGI or HLMI (as defined in 'What capabilities could constitute TAI?'). For example, they need not perform especially well – or perhaps even *at all* – in areas such as strategic leadership, policy-making, advanced planning and scheduling, financial decision-making and budget management, customer services, stakeholder management, sales, entertainment, or even any scientific research that falls outside of the relatively circumscribed domain of AI R&D.[62]

[60] In the literature, it is commonly imagined that the AI R&D field would begin by deploying AI systems that *initially* perform a fairly low proportion of the tasks typically performed by researchers/engineers (say, 40%) and humans therefore remain heavily involved – but through the effects of the direct feedback loops which emerge here, capabilities progress is subsequently fast, and we soon end up with systems which can perform near-100% of these tasks (and the role of humans diminishes accordingly). It therefore doesn't matter too much (for our purposes) what we think the *exact* threshold for task performance should be for an AI system to count as an 'automated worker': if we start off by deploying an AI system that is slightly below this threshold, we might quickly end up with systems that exceed it.

Shortly, I spell out an argument for recursive improvement via AI R&D automation in more detail. Though I use the framing of 'automated workers', I don't specify whether there is already near-100% automation of AI R&D at the beginning of the story, or if we instead begin with much lower levels of automation that increase over time.

[61] I also think it's more plausible in this case than under most other framings of direct recursive improvement which are comparably strong. However, I do not get into direct comparisons in this report.

[62] In 'Other objections', I'll also note that automated AI R&D workers need not exhibit the same kind/degree of *generality* as is necessary under (most conceptions of) TAI.

It is still possible that a wide range of other AI capabilities could end up being developed *in tandem* with those which are most relevant for AI R&D, such that automated AI R&D workers happen to arrive at the same time as some form of TAI. But there are also conceivable stories in which we *do* get automated AI R&D workers in the pre-transformative era, and these workers actively drive the development of transformative systems. These are the stories I will keep in mind for the rest of this chapter.

A direct recursive improvement story

- Equipped with this framing, we can spell out an argument for direct recursive improvement in the following way[63]:
- At some point before TAI arrives, we get AI systems that can act as drop-in replacements for researchers and engineers. I variably call these 'automated workers' or 'automated AI R&D workers'.
- We will then be able to run large numbers of these AI systems (perhaps many millions, according to Aschenbrenner, given the availability of inference compute at this point) and set them to automating *all or most* of AI R&D.
    - Human researchers and engineers may or may not continue their own work alongside these automated workers; this detail is not critical to the story.
- This would vastly increase both the number and quality of workers in the AI R&D field. Not only would there be a much bigger workforce, but these automated workers would also be able to run much faster than human workers (say, 10x faster), and could be deployed for 24 hours a day without needing breaks or losing focus.
- With huge increases to human-equivalent hours spent on AI R&D, there would inevitably be improvements in the field, likely driven by progress on the components of the AI Triad.
    - Some people who tell this story focus primarily on the effects of software improvements, but the automation of AI R&D could also improve hardware and therefore speed up the scaling of physical compute.
- This could introduce an 'AI R&D type' direct feedback loop. The first generation of automated R&D workers could make improvements to a second generation of models. This generation would therefore be even more capable than the first on a range of tasks – including, but not limited to, those required for AI R&D. It would thus be even better equipped to make improvements to subsequent generations of models than its predecessors were.[64]

Shortly, I will outline a few different options for what the step-by-step trajectory of generational improvements might actually look like in worlds like

63 Importantly, Aschenbrenner, Davidson, and Drexler provide models/arguments for *different variants* of the story below. There are some differences between their prominent characterisations of recursive improvement and what I describe in the context of timelines to TAI. For example, Aschenbrenner introduces recursive improvement specifically as a pathway taking us from AGI to ASI, while I present it as a pathway from weaker-than-transformative systems to the first transformative systems. Importantly, the same arguments can be (and have been!) extended to support both of these stories.

64 Harking back to the ambiguity of my initial characterisation of automated workers, and an earlier footnote: one could imagine beginning this story with a generation of AI systems which cannot yet perform a very high proportion of R&D tasks (but might still reasonably be labelled as 'replacing' human workers in some meaningful way). Over a series of generational improvements, these systems would become increasingly capable across a range of tasks, with decreasing need for human involvement. In that case, we'd gradually approach ~100% automation, even if we didn't start there. And as the level of automation in the field increased, so too would the strength of the direct feedback loop in play. A gradual automation story of this kind has been explicitly modelled by Tom Davidson through his compute-centric framework for takeoff speeds.

this. But first, I note an alternative way of thinking about feedback loops in AI R&D.

Other framings of AI R&D automation

The above is a popular way of telling the AI R&D automation story, and the one I'll favour. However, it's not the only option.

Plausibly, we could see some recursive dynamics emerge without having any individual system that can fully *or even nearly* automate the role of an AI R&D worker. Direct feedback loops of a similar (albeit more restricted) kind could be introduced by more narrowly capable systems which fall *far* short of being recognisable as a 'drop-in replacement' for human researchers or engineers. Examples include:

- AIs which **generate synthetic data** for training models (but do not contribute directly to areas of AI R&D outside of enhancing the data pipeline)
- AI **coding assistants** which improve human productivity in software development (but do not autonomously make software improvements)
- AIs which support human workers by **generating ideas** for future AI R&D (but cannot execute on these ideas)

Although an AI system of any of these kinds could introduce some recursive behaviour within the field of AI R&D, this would likely be *much more limited* than the recursive behaviour enabled by systems which can act as drop-in replacements to human researchers and engineers. Unlike the automated worker case, the recursive improvements enabled by the individual AI R&D services and tools described above would be scoped to specific narrow tasks and/or still depend heavily on human involvement. Overall, achievements in the AI R&D field would still be driven in large part by human researchers and engineers.

Accordingly, I expect these direct feedback loops to (individually) have weaker effects[65] on capabilities progress than those characterised under the automated worker framing, which are brought about by more broadly capable AI R&D systems acting with higher degrees of autonomy.

However, it's possible that a *combination* of these weaker feedback loops could (in some sense) effectively add up to the deployment of automated AI R&D workers. That is to say: the combined effects of a fleet of varied, narrowly capable AI systems, each supporting capabilities improvements in their own limited ways, might closely *approximate* the effects of introducing 'automated workers' into AI R&D.

To get a sense of what this might look like, imagine a world in which humans have access to **a diverse suite of productivity-boosting AI tools and automated services** for AI R&D, such as AI coding assistants and idea generators (henceforth '**suite of services**'). Suppose also that these narrow systems can jointly cover *most* aspects of AI R&D.

[65] What I mean by the 'strength' of the effects of different feedback loops will become clearer in the following subsection on the possible trajectories of recursive improvement.

In this world, the actual role of a human researcher or engineer is now greatly reduced. There are lots of ways we could flesh out the details here, each affording slightly different roles to this human, but one very plausible story is that she now plays a *supervisory role* in AI R&D – issuing instructions, checking each individual system's outputs at defined points in the R&D process, and selecting and synthesising the outputs from different systems in useful ways.

As in the automated worker case, each individual system comprising this suite runs faster and for more hours of the day than humans can. Overall, productivity in AI R&D is dramatically increased.

We could therefore tell a variant of the direct recursive improvement story above which refers to a suite of AI R&D services instead of automated workers. In fact, this framing seems more in line with how Drexler initially imagined the process of AI R&D automation.

Comparing stories of AI R&D automation

Adopting the above idea of a 'suite of services' over the 'automated workers' framing of direct recursive improvement would slightly, but not significantly, alter the arguments of this chapter.

Importantly, we have already established that the potential *scope* of contributions to AI R&D is similar in both cases.

Now, the limited autonomy of the individual AI systems in the 'suite of services' case might restrict the speed of contributions to AI R&D: even if the systems involved are performing their tasks extremely fast, capabilities improvements cannot occur any faster than the pace at which humans can, say, synthesise the outputs from these different systems.

This difference is unlikely to be dramatic, since even the 'automated worker' framing admits *some* role of humans in supporting AI R&D tasks.

It does seem, however, that humans might play *a more crucial role* in the 'suite of services' story than in the 'automated worker' story – at least as these stories are currently fleshed out. In the 'suite of services' case, each model output might need to go through human supervisors at a few critical steps of the given R&D workflow (e.g. the prompting, checking, selecting, and synthesising steps).[66] By contrast, this does not feel like a natural part of the 'automated workers' story. In fact, since the *selecting* and *synthesising* steps seem like non-trivial aspects of most R&D workflows, requiring humans to perform these steps might count as *significant reliance* on humans, which I've stipulated would not happen in the automated workers case. So (tentatively): one might expect humans present more of a bottleneck to capabilities progress in the 'suite of services' story than in the 'automated worker' story.[67]

As a result, some details of the debate over *how fast* and *how sustained* the progress from direct recursive improvement will be (which is the focus of 'Why recursive improvement might not produce a short TAI timeline') do depend on the chosen framing. I'll flag this up briefly in cases where it seems especially useful to the reader to be aware of, but I won't dwell too much on the specific

[66] Of course, there is far *less* human involvement in this story than in a case where *only one* aspect of AI R&D (say, idea generation) gets automated. But humans are still crucially involved at certain supervisory stages of the R&D process, even if there's no particular area of AI R&D work that some specialised AI systems cannot contribute to.

[67] I think this result may be more of a feature of the way I have chosen to tell these stories than anything fundamental to the type of systems I'm describing. It's also very sensitive to questions like "what percentage of AI R&D is actually automated" in each case, which I haven't answered. I therefore try not to *overemphasise* the effects of any perceived differences in the level of AI autonomy under these two framings of direct recursive improvement.

differences.

There might also be some meaningful difference in the likelihood of these two different forms of direct recursive improvement emerging within the next few years: the 'suite of services' might be lower-hanging fruit for developers, given that it doesn't rely on developing individual systems which are competent on such a wide range of tasks, and also might not require these systems to complete those tasks as independently as they would in the 'automated worker' case. However, both options do seem *plausible* in the near term, as will become evident in 'Other objections'.

For the rest of this chapter, I'll mainly go ahead with the 'automated worker' framing. This is partly for its convenience and simplicity, partly due to its popularity in the recent literature, and partly because it aligns closely with what some labs are actually *trying* to develop (i.e. broadly capable AI agents; a good example in the R&D context is Sakana AI's 'AI Scientist'). However, the existence of a variant of this automation story involving AI systems which are potentially even easier to develop *and* could have roughly similar effects on capabilities progress is a point in favour of those who believe in short timelines via direct recursive improvement.

In this subsection, I have introduced a lot of new terminology that we will need to keep in mind in the second half of this chapter. We also need some additional, more neutral language for talking about *any* setup of AI systems that can automate significant parts of AI R&D without distinguishing sharply between the two different framings highlighted in this section. I therefore supply a full glossary of key terms below.

Glossary of key terminology for direct recursive improvement

Automated (AI R&D) workers: AI systems which can individually perform *all or most* of the work done by a human AI researcher or engineer (to the same standard or higher, end-to-end, and without significant reliance on humans). These systems can act as 'drop-in replacements' for human AI R&D workers.

Suite of (AI R&D) services: A combination of AI services and/or tools for AI R&D which can be jointly deployed with (roughly) similar overall effects to the 'automated workers' described above.

R&D AIs: Refers broadly to AI systems (or networks of systems) which contribute significantly to AI R&D, without singling out a particular interpretation of how these contributions are made. This term encompasses both 'automated workers' and 'suites of services'.

Automated (AI R&D) workforce: This could be comprised of a variety of different R&D AIs, or many copies of one R&D AI system.

'AI R&D type' direct feedback loops: Positive feedback loops mediated by R&D AIs (for example, automated workers/suites of services) which contribute to the AI R&D field, and become more capable of making subsequent contributions in the process.

Some of these feedback loops are more 'restricted' than others, and therefore likely to have weaker effects. For example, a single AI tool which assists human workers on a specific AI R&D task would be more restricted in its contributions to capabilities improvements than a fully automated worker or a diverse suite of services. In this chapter, I focus on feedback loops in AI R&D which are broader in scope and less reliant on human involvement.

Direct recursive improvement (DRI): In this report, I use this term to refer to an iterative process of capabilities improvements driven by 'AI R&D type' direct feedback loops (of the less restricted kind stipulated above). A variety of different processes could also warrant this label, but these alternatives are not discussed in this chapter.

Figure 2.2: Glossary of key terms associated with 'direct recursive improvement'.

Possible trajectories of capabilities improvements via DRI

Regardless of the exact formulation we choose, stories of direct recursive improvement are (as noted) *structurally similar*: we cross some capabilities threshold beyond which some AI systems can take over some crucial aspects of AI development and increase AI capabilities iteratively. At each step, the systems in question are better able to make these improvements. But what does the resulting step-by-step trajectory of capabilities improvements actually look like? Below, I briefly outline several possibilities.

A sustained period of acceleration? People who believe direct recursive improvement will produce a short timeline sometimes argue that each (time-equivalent) step will represent a greater jump in AI capabilities than the previous one – or, equivalently, the same magnitude of capabilities jump will take an increasingly small amount of time to achieve. This would characterise "a sustained period of accelerating growth" in AI capabilities. Several influential thinkers, such as Nick Bostrom and David Chalmers, have suggested that this could underpin rapid exponential or even superexponential trajectories of AI progress.

Some more recent support for these ideas can be found in quantitative research pieces by Epoch and by Tom Davidson, which are discussed in the below subsection entitled 'Will capabilities progress accelerate?'.

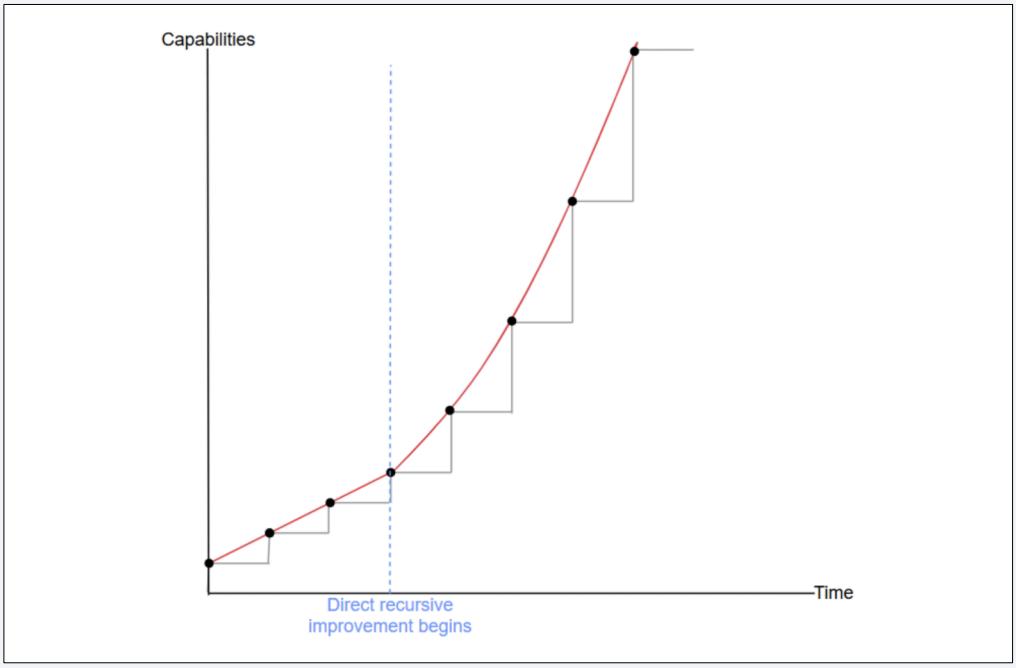


Figure 2.3: An exponential trajectory of capabilities steps.

However, a runaway trajectory of progress like this does not follow immediately from the mere assumption that AI systems *get better at making capabilities improvements* at each step of the direct recursive improvement process. Indeed:

- If, at the same time, *capabilities improvements get more difficult* at each step, then AI systems possessing better capability-improving abilities may

not actually translate into making bigger jumps in capabilities.

- If there are *constraints on capabilities improvements at each step*, then increasing capability-improving abilities cannot lead to arbitrarily large jumps in capabilities.

These points will be explored further in the counterarguments section of this chapter.

A one-time speedup of progress? Even if direct recursive improvement does not underpin *exponential* or *superexponential* growth in AI capabilities, it could still have a significant effect on the trajectory of progress.

For example, if each capabilities jump under direct recursive improvement is at least *greater* in size than the equivalent jump that human R&D workers alone would have produced in the same time, then we'll see faster growth in capabilities than we would have otherwise. This may not characterise a sustained period of acceleration, but *could*[68] still correspond to a sudden – and possibly dramatic – step change in growth rate.

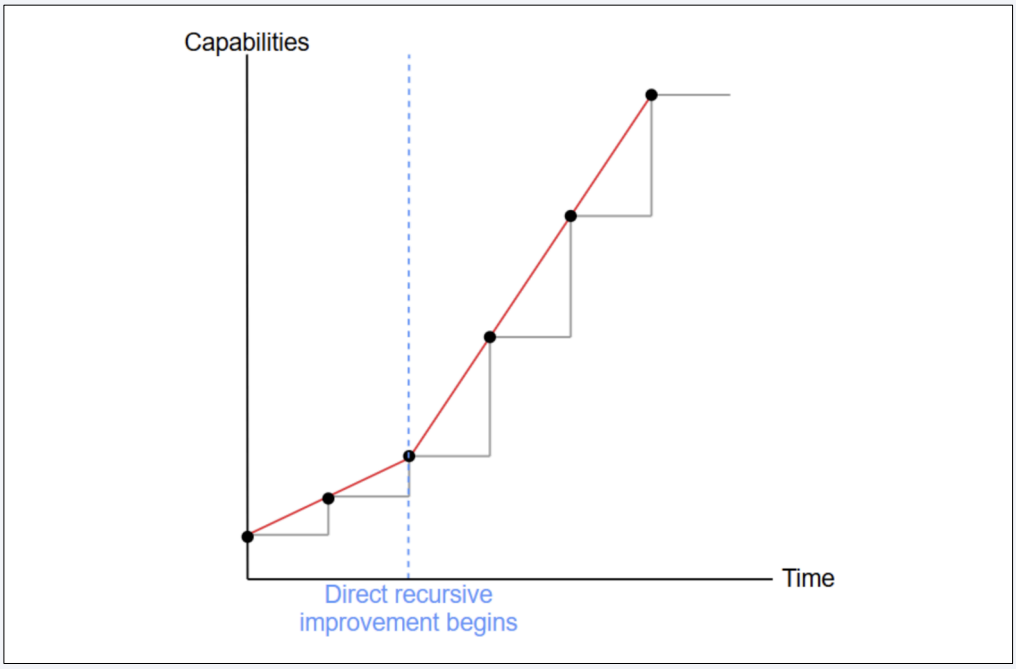


Figure 2.4: A step change in capabilities growth rate.

If this is what happens at first, and it's a dramatic enough step change, it might not even matter whether something similar applies for every subsequent jump: perhaps the first step on the ladder will already be enough to get us to TAI.

[68] Not necessarily, though. The magnitude of capabilities jump that human R&D workers could produce over a fixed period of time could itself be decreasing, or about to plateau, at the time at which direct recursive improvement kicks in. In that case, even if automation enables us to achieve *more* than human R&D workers would have achieved alone, we might just see a continuation (or even a decrease) to current rates of AI progress. I will gesture to these possibilities under the 'Slower trajectories' subheading.

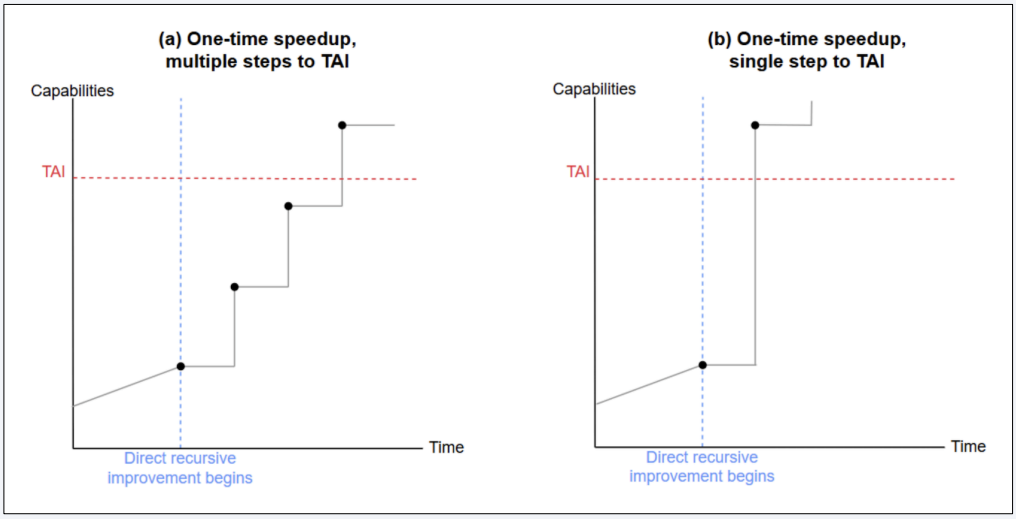


Figure 2.5: Multiple steps vs single step to TAI.

In Leopold Aschenbrenner's words:

> "Yes, recursive self-improvement, but no sci-fi required; … *[the automation of AI R&D] would need only to accelerate the existing trendlines of algorithmic progress*".

This idea will be revisited in more detail in 'Who wins the tug of war?'

Subexponential trajectories? Some possible trajectories of direct recursive improvement don't involve *any* increase in the pace of capabilities progress at all. For example:

- *Current rates of capabilities progress could more or less continue as they are*, if serious bottlenecks to capabilities progress (e.g. the factors which could affect compute scaling, as highlighted in the previous chapter) have emerged by this point in time, and the positive feedback loops underpinning recursive improvement are merely strong enough to compensate for them.
- *Current rates of capabilities progress could still decline*, if the feedback loops underpinning recursive improvement are too weak to even compensate for the increasing effect of bottlenecks to progress.

The relationship between recursive improvement and the potential bottlenecks to compute scaling is discussed in the section of this chapter titled 'Overcoming scaling bottlenecks via recursive improvement'.

Note: neither of these subexponential trajectories of capabilities progress are *incompatible* with a short timeline – but the case for short TAI timelines is clearly stronger if we do expect capabilities progress to accelerate.

Will capabilities progress accelerate? Quantifying the effects of direct recursive improvement

Several researchers have attempted to quantify the potential effects of AI R&D automation on the trajectory and timeline to TAI (accounting for *at least some* of the possible constraints to progress and diminishing returns to effort that I

explore later in this chapter). Overall, these sources provide some support for the claim that direct recursive improvement would cause capabilities progress to accelerate.

I highlight some relevant examples below, alongside their key findings. I encourage the interested reader to explore these further herself.

Epoch AI. Epoch's *Estimating idea production: a methodological survey* (and the corresponding blog post) investigates the returns to automating software AI R&D, taking into consideration the effects of diminishing returns from effort.

The Epoch team examines the rate of returns in each of several different domains of software. If this rate exceeds 1 in some domain, this is understood to be "consistent with a proportional increase in research input resulting in greater-than-proportional software improvement" within that domain. In other words: a rate of returns exceeding 1 within some domain is consistent with *a trajectory of continuously accelerating progress* within that domain.

Across the software domains studied by Epoch, the median rate of returns is in the range 0.8-3.5. Several (but not all) of these are found to have a median exceeding 1.

According to the authors, this evidence "hint[s] at the possibility" that automating AI R&D could lead to accelerating progress in capabilities *overall*, but is far from conclusive on this issue.

Tom Davidson. In his work on compute-centric frameworks for AI timelines, Tom Davidson's 'Full Takeoff Model' (FTM) incorporates the effects of gradual AI R&D automation on capabilities progress, in comparison to a 'toy model' which does not.[69] Like Epoch, he also accounts for some effect of diminishing returns here.

According to Davidson's model, factoring in the dynamics of AI R&D automation "causes growth of the three components [which determine effective FLOP] to accelerate continuously. By the time we reach full automation of cognitive labour (AGI), they can double in months or faster". Based on his modelling assumptions, this leads to superexponential growth in effective FLOP, which underpins a period of accelerating progress in capabilities.

This has a significant impact on the timeline to TAI. The inclusion of AI R&D automation dynamics results in a 2.5x reduction in the time between AIs being able to perform *20% of cognitive labour* to AGI (operationalised as AI being able to perform *100% of labour*).

In a similar vein, a forthcoming article by Tom Davidson models the effects of *fully* automating AI R&D on software progress. He estimates that this would initially speed up current rates of progress 2-30 times (and that significant intermediate speedups from partial automation are also possible). He also argues that despite diminishing returns, it is still more likely than not that we'd see an *initial* period of continuously accelerating progress (though the rate of improvements would *eventually* slow down).

[69] Note that Davidson models direct recursive improvement as a gradual process in which levels of automation in the AI R&D field increase over time. There is no specific threshold he highlights in this story at which we can say 'direct recursive improvement has begun'.

The route to transformative AI

Regardless of the *exact* trajectory of capabilities improvements we expect to unfold here, the basic argument for the rapid emergence of TAI looks more or less the same. Once AI capabilities reach some important threshold enabling significant AI R&D automation, progress could get very fast – perhaps increasingly fast, but not necessarily. Either way, TAI could follow shortly thereafter.

As mentioned earlier, the implications of this argument on the timeline to TAI depends on at least four variables: how fast the resulting growth in capabilities actually is, how long the growth period is sustained for, the time at which this growth period begins, and how far away from TAI-level capabilities we are at that time.

But, if we do reach this threshold within the next few years, and the growth mode is sufficiently fast, then TAI might arrive by 2035. Indeed, Tom Davidson's Full Takeoff Model, which incorporates the effects of AI R&D automation in its timeline projections, places 20-40% weight on the arrival of TAI by 2035 – a considerable chance of a short TAI timeline being realised. His inclusion of recursive improvement dynamics is likely one of the reasons his model places more weight on short timelines than Cotra's (the differences in their predictions was highlighted in Table 1.1).[70]

Further reading on direct recursive improvement:

- *Chapter 4 of Nick Bostrom - Superintelligence*
- *Eliezer Yudkowsky - Intelligence Explosion Microeconomics*
- *David Chalmers - The Singularity: a philosophical analysis*
- *I. J. Good - Speculations Concerning the First Ultraintelligent Machine*
- *Tom Davidson - What a compute centric framework says about AI takeoff (Summary post; full report)*
- *Tom Davidson - Forthcoming article on the size and speed of an 'intelligence explosion'*
- *Epoch - Do the Returns to Software R&D Point Towards a Singularity? (and the full paper)*
- *Chapter II of Leopold Aschenbrenner - Situational Awareness*
- *Chapter 1 of Eric Drexler - Reframing Superintelligence*

Indirect recursive improvement

Some feedback loops that are relevant to AI capabilities progress are not directly mediated by AI systems. Instead, they concern the cyclical progress of a much broader societal system, which indirectly brings about improvements in the specific field of AI R&D. AI systems themselves may feed into this process to some extent, but unlike in the direct case, they are not the key *lever* for recursive improvement here.

[70] In an accompanying blog post, Davidson notes that one "major change [from Cotra's model] is that … [he] model[s] the effects of AI systems automating economic tasks – and, crucially, tasks in hardware and software R&D – prior to AGI".

These 'indirect' forms of recursive improvement are not what people are typically referring to when they talk about recursive improvement in the context of AI timelines. However, their influence on the dynamics of AI progress is significant, and worth briefly noting.

These feedback loops include:

- **Economic feedback loops.** AI systems could contribute significantly to the economy. Indeed, McKinsey has noted that "generative AI has the potential to generate $2.6 trillion to $4.4 trillion in value [annually] across industries". AI labs will likely receive some fraction of this money by charging for API access. Large amounts of money could then be reinvested into further AI capabilities improvement, potentially bringing in even greater economic returns, which could once again be reinvested – and so the cycle would continue. In addition, economic competition may cyclically increase the impetus for AI labs to make advances in system capabilities, as each lab continuously strives to outperform its competitors.
- **Scientific feedback loops.** Regardless of whether we see recursive improvement mediated by AI systems themselves, the broader system of scientific development will still be experiencing feedback loops of its own, and these will in turn influence the development of AI. New scientific insights help us develop enhanced tools and methods, thereby equipping us to make insights more easily in the future. This may support progress in areas relevant to AI capabilities, such as hardware development.
- **Political feedback loops.** These may be especially strong in the context of international security and geopolitical race dynamics. When one actor makes advances in AI development, others are motivated to do the same, which in turn provides even further motivation to the original actor to stay ahead.

The route to transformative AI

It should be noted that indirect feedback loops like these are *already* happening and, as such, are already affecting AI progress to some extent. AI labs are already attracting capital[71] to invest into AI R&D, allowing bigger and better models to be built; we're beginning to see leading AI labs race against each other to advance system capabilities[72]; broad scientific progress is providing us with better tools and hardware for building more capable AIs; and competitive political pressures seem to be increasingly driving state investment into AI[73]; these may be partial explanations for recent growth in AI capabilities. If these feedback loops gain more traction, they will likely speed up AI progress.

In fact, it's not clear how AI capabilities could develop very rapidly over the next ten years *without* increased contribution from some of these feedback loops. The other mechanisms for fast AI progress I have described – both

[71] This doesn't *strictly* take the form of the economic feedback loop characterised above, which is based on labs reinvesting profits from their AI systems. Leading AI labs are not yet making *profits* in this way – but, importantly, the market is rewarding them via a different route, constituting another kind of economic feedback loop.

[72] See, for example, the recent interest in AI leaderboards.

[73] See, for example: indications of competition between the US and China in Cullen O'Keefe's work on 'Chips for Peace' and later chapters of Aschenbrenner's *Situational Awareness*; as well as in AI, Global Governance, and Digital Sovereignty by Swati Srivastava and Justin Bullock, which characterises AI development as a playing field through which states compete for power.

compute scaling and direct recursive improvement – rely on the availability of increasing amounts of investment to sustain them, the pace of relevant scientific advancements generally keeping up, and increasing motivations to build more capable systems.

This observation is (probably) not unique to the AI context; feedback loops of this kind have also been key enablers of fast progress in other technological fields. Some of the most notable historic cases of rapid technological development were seen during the Cold War, in which races for progress in space travel and nuclear arms technology yielded major advancements over a short period of time. It seems unlikely that humans would have travelled to the moon or developed such powerful nuclear arms *so quickly* had there not been strong feedback loops in play with respect to geopolitical competition and investment. (Indeed, progress in both space travel and nuclear arms seemed to stall somewhat when these feedback loops dissipated.)

Although some degree of indirect recursive improvement is likely to be necessary for a short timeline to TAI, this may not be a sufficiently convincing story for short timelines on its own. A story in which direct recursive improvement is *supported by* relevant indirect feedback loops has more argumentative weight: the interplay of direct and indirect recursive dynamics would likely yield much faster capabilities progress than the continuation of indirect feedback loops *alone*.

Put another way: indirect recursive improvement is already happening, and will likely continue. It is the advent of direct recursive improvement dynamics in AI – dynamics which we haven't seen a clear equivalent to before in the historic development of AI – that could be a real game changer for the entire course of AI progress. The introduction of direct recursive improvement into our story of AI progress is something that could make our case for short timelines substantially more convincing.

In later sections of this chapter, when I consider counterarguments to recursive improvement as a pathway to short timelines, I will therefore be focused on *direct recursive improvement*. That is, I will simply assume that the indirect feedback loops relevant to AI progress are continuing in the background, and gaining traction to whatever extent is necessary to make a short TAI timeline feasible. The evaluative aspects of these chapters will concern *how direct recursive improvement* (if adequately supported by these other necessary feedback loops) *could succeed or fail in bringing about a short TAI timeline*. Nonetheless, some of the arguments considered may be broadly applicable to less direct forms of recursive improvement.

Overcoming scaling bottlenecks via recursive improvement

Before I turn to examine specific counterarguments to recursive improvement as a pathway to short TAI timelines, it's worth reflecting on the relationship

between recursive improvement scenarios and the arguments of the previous chapter.

In the previous chapter, I outlined how AI capabilities growth based on compute scaling could possibly be bottlenecked in a number of ways.[74] The bottlenecks identified there could have a slowing or stalling effect on AI progress over the next decade.

However, if fast recursive improvement dynamics kick in and have sufficient impact on the bottlenecked areas, some of these challenges could plausibly be overcome. (By this, I don't mean to say that such challenges will not emerge for, or have any impact on, the pathway to TAI in a recursive improvement scenario. But if they do emerge in this scenario, some of their effects might quickly be counteracted, such that TAI could plausibly still emerge within the next ten years.)

Some of these bottlenecks might be most plausibly overcome in scenarios featuring direct recursive improvement (DRI), while others may be counteracted sufficiently through indirect recursive improvement (IRI) dynamics alone.[75] For example:

- **Algorithms and the current paradigm.** If algorithms become the primary bottleneck for capabilities improvements, recursively improving generations of 'R&D AIs' (defined in Figure 2.2) might become increasingly able to find solutions to the current limitations of LLMs – whether this requires developing a completely new paradigm, or making efficiency improvements and providing new tools/scaffolding to the existing one. This is most plausible in DRI scenarios.
- **Data, environment, and virtual embodiment.** R&D AIs might become increasingly well-placed to tackle challenges concerning data and training, as they improve their ability to generate vast quantities of synthetic data (or to search for the highest quality data available, through spending more inference compute). They could also become iteratively better at designing training environments for successor systems (and if necessary, at some stage, even be able to create a sufficiently rich simulation of the world to train 'virtually embodied' AI systems). *This is most plausible in DRI scenarios*.
- **Hardware.** An AI R&D workforce with increasing capacity and skill could also speed up progress in hardware, possibly reaching a point where it can overcome bottlenecks on the side of hardware efficiency and compute availability. *This is most plausible in DRI scenarios in which R&D AIs aren't exclusively deployed to work on software*.
- **Investment.** Fast economic feedback loops could break any bottlenecks related to money. Lack of motivation is also unlikely in this scenario, as rapid economic returns from AI development would provide clear incentives for investment.[76] *This is plausible in certain IRI scenarios, even without the emergence of DRI dynamics.*

[74] Specifically, I identified constraints for compute growth, as well as bottlenecks that could slow down capabilities improvements even if compute growth is very fast.

[75] Note that what I'm calling 'DRI scenarios' in this section are scenarios which feature DRI as well *as any other feedback loops necessary to support DRI*. According to the reflections of the previous section, this means that DRI scenarios do not preclude – and in fact, they probably require – some degree of IRI. When I use the term 'IRI scenario', I'm remaining silent on whether DRI dynamics have emerged in that scenario, unless specified otherwise. This term will be used when the emergence of DRI is not necessary for the argument at hand.

[76] It's less clear how some of the other constraints for AI progress described in that chapter, such as those concerning the supply chain, might be overcome through some form of recursive improvement.

Positive feedback loops (if sufficiently strong) could thus serve as a powerful mechanism for capabilities progress, improving the field's abilities to overcome challenges on the path to developing TAI and the speed at which it does so.

They might be especially powerful in unblocking and reinforcing capability gains from compute-based scaling. Indeed, popular stories of short timelines often feature a *combination* of compute scaling and recursive improvement enabling rapid AI progress synergistically, rather than progress being driven by any one mechanism in isolation. Such synergies will be captured in the short TAI timeline scenarios characterised in Chapter 3.

It is not clear, however, that any such feedback loops and associated recursive improvement dynamics will *actually* gain enough momentum in the next ten years to quickly overcome the challenges described above. (And even if they do, it's not clear whether their effects would be strong enough to accelerate capabilities progress, or just sustain current trendlines in the face of bottlenecks.[77])

In fact, there are reasons to think that recursive improvement steps would actually not be very fast, or would quickly fizzle out. This is the primary subject of the next section.

Why recursive improvement might not produce a short TAI timeline

> NOTE
>
> Throughout this section, I focus on DRI scenarios driven by 'automated workers' (though some arguments will also be applicable to scenarios exclusively featuring IRI, and some will be applicable to DRI scenarios driven by a 'suite of services'). Since I lean on the terminology I introduced in the first half of this chapter, it may be helpful for the reader to refer back to the glossary provided in Figure 2.2.

Points of disagreement. Will recursive improvement produce a short TAI timeline? Let's imagine a sceptic and believer taking opposing stances on this question, with several possible areas of contention between them.

In this section, I will basically grant to the believer that DRI dynamics *will emerge* at some point in the future. Specifically, I'll assume that:

- The AI field will eventually reach a point at which some AI systems are highly skilled at important parts of AI R&D. (For framing purposes, I imagine that these systems will take the form of 'automated workers', but since we've noted that *being highly skilled at important parts of AI R&D* could be variously interpreted, I'll often use the broader term 'R&D AIs' here.)

[77] While short TAI timelines are possible in either case, their likelihood is higher in the instance where progress rates are accelerated, rather than merely sustained, in the next ten years.

- After this point, these R&D AIs will begin driving capabilities improvements, through an iterative process exhibiting some degree of positive feedback: with each capabilities step that then occurs, the field is, overall, in an improved position to make future capabilities improvements. I'll describe this situation as a *period of (direct) recursive improvement*.

A sceptic could, of course, question the assumption that a period of (direct) recursive improvement will ever happen. However, what I think is most interesting to the timeline debate here is that *even if we do grant this assumption to the believer*, it's not immediate that this would imply a short TAI timeline.

As noted earlier, supposing a period of recursive improvement in AI capabilities progress occurs, its implications on the timeline to TAI are determined by the answers to the four following questions:

(1) How fast is the capabilities growth resulting from the feedback loops at play? (i.e. what is the shape of the trajectory of recursive improvements?)

(2) How long is this period sustained for?

(3) When does this period begin?

(4) How far away is TAI when this period begins?

A sceptic could take up any of these four questions as an angle for arguing that if a recursive improvement period happens, it will still fail to produce a short TAI timeline. That is, she could adopt some combination of counterarguments along the following lines:

(S1) The relevant feedback loops *might just not lead to very fast improvements in AI capabilities*;

(S2) The *resulting capabilities improvements might plateau or stall* at some point before TAI arrives (rather than being sustained for enough time to bring about TAI);

(S3) The relevant feedback loops *will not kick in soon enough* for TAI to emerge within the next ten years;

(S4) When the period of recursive improvement begins, *the threshold for TAI will be so far away* that it will still not feasibly be met within the next ten years.

A plan for the rest of this chapter. I'm most interested in arguments around the first two points, (S1) and (S2) above.

These two claims aren't cleanly distinguished from each other in the literature, since the basic arguments underlying them are often similar; I therefore don't typically distinguish them explicitly in this section.

My approach will be as follows: I'll suppose that some recursive improvement period begins in which AI systems begin to make step-by-step improvements to AI capabilities. I'll then examine *three separate ways* the sceptic could argue

that these capabilities steps will be quite small or slow to make (as in S1) or might even plateau before TAI arrives (as in S2).

Put simply, these three arguments are:

- Each capabilities step will be harder to make than the believer imagines, due to certain situational constraints and requirements. ('Each capabilities improvement step is more difficult than expected')
- Each capabilities step will be harder to make than the previous one, due to the increasing effect of bottlenecks and diminishing returns on effort. ('Capabilities improvements get more difficult at each step')
- Even if inputs to AI capabilities improve in big or fast steps, AI capabilities themselves might only improve in small or slow steps. ('The relationship between accessible input improvements and capabilities improvements is sublinear')

I will advance and examine each of these three arguments individually over the next three sections. My main reference points for these arguments within the literature are notable works by François Chollet and David Thorstad. However, much of the exposition – especially for the first argument – is my own.

The relationship between these three counterarguments and the key questions highlighted earlier is visualised below.

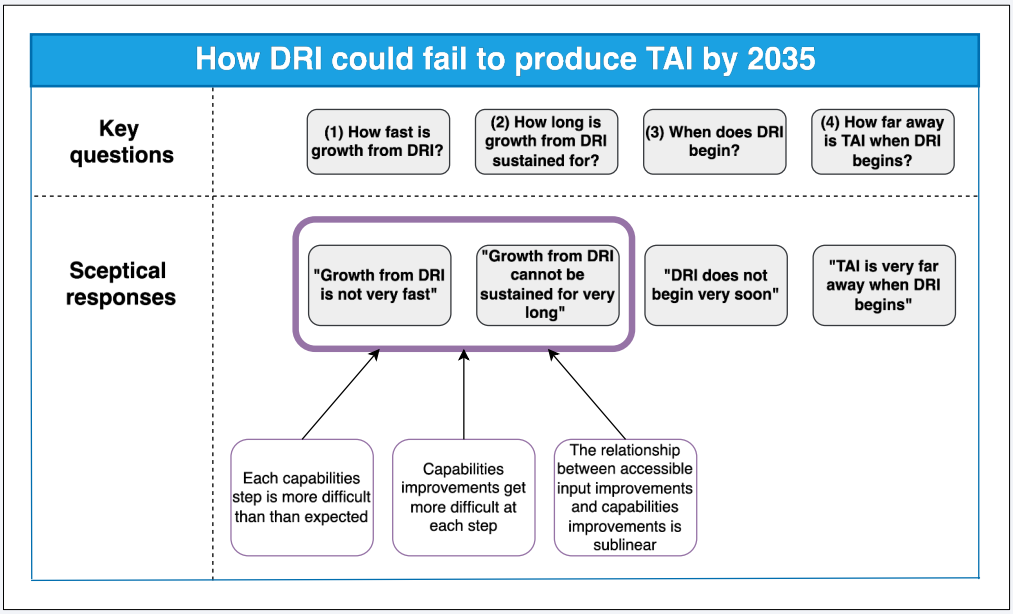


Figure 2.6: How direct recursive improvement could fail to produce TAI by 2035. This chapter focuses mainly on key questions (1) and (2), though some attention is paid to the other two questions at the end of the chapter.

In 'Who wins the tug of war?', I'll consider the implications of the debates so far on the plausibility of short timelines. Through this, it will become clear that an 'explosive' trajectory of capabilities improvements is not an essential part of the story for short timelines via DRI.

Finally, in 'Other objections', I'll briefly illustrate how the sceptic might also contest short timelines by advancing claims (S3) and (S4) on the above list.

Each capabilities improvement step is more difficult than expected

NOTE

I'll advance this counterargument with specific reference to the '**automated workers**' framing of DRI. It's possible to make an analogous argument with reference to a 'suite of services' for AI R&D, with some meaningful adjustments. I do not walk through the details of the necessary adjustments.

The basic objection

Many stories of AI progress featuring DRI make the following simple assumption: that at each step of the recursive process, certain AI systems will be able to *use* their increased capabilities to make further improvements to the capabilities of AI systems.

For this to result in a short TAI timeline, the believer must also maintain that it will be *relatively straightforward and quick* for these AI systems to make each of these improvement steps on the pathway to TAI.

One possible objection to the prospect of short timelines via DRI is that making such capabilities improvement steps might be *much harder for AI systems to achieve* than the believer has imagined – either in general, or at least some of the time.[78] If so, then DRI-mediated capabilities improvements might be slower or harder to sustain than would be required for the timeline to TAI to be short.

Crucially, this objection is not a denial of the simple assumption that some AI systems *will be able* to make further improvements to AI capabilities. That is, the sceptic here is not claiming that direct feedback loops will fail to emerge or get off the ground. Instead, she is better seen as claiming that the positive feedback from one step of the recursive process to the next is just not as strong as the believer has argued, and as such, it will be *harder* to make these improvement steps than the believer thinks.

Or, put differently, the objection is as follows: simply having higher levels of AI system capabilities will not straightforwardly or consistently empower AI systems to make significant further improvements on those capabilities, because the ability to make capabilities improvements is not just a function of existing capabilities levels. Instead, it hinges upon certain other factors which may have serious constraining effects on progress.

What *do* capabilities improvements hinge on? Some sceptics of recursive improvement, such as François Chollet, have emphasised *situational constraints on AI systems* as key challenges for capabilities improvements via DRI. Here, I advance a Chollet-like argument: I draw heavily from his article

[78] For simplicity in this section, I frame this objection in terms of the stronger claim here: that capabilities steps will be *generally* hard to make. But even if just one particular step is extremely difficult, and all others are straightforward, this one-time challenge could still, in theory, delay progress sufficiently to rule out a short timeline.

The implausibility of intelligence explosion and lightly incorporate some ideas from his more recent work, but provide my own gloss and analysis throughout in order to engage more closely with the specific formulation of direct recursive improvement (with automated workers) that I gave earlier in this chapter.

I will shortly motivate and describe the situational constraints highlighted by Chollet, and how difficult they might be for the AI field to contend with. But first, I will begin with some notes on the applicability of this kind of argument to different forms of DRI, and on its potential implications for TAI timelines.

Narrowing our focus to automated workers

This kind of argument could be advanced in different ways, corresponding to the many different senses in which AI systems might be described as 'making AI capabilities improvements'. In this chapter, I've already narrowed our attention to scenarios in which AI systems make capabilities improvements *via contributing to AI R&D*, but this admits various interpretations (including the two framings 'I introduced in the 'Direct recursive improvement' section).

In this section, I'll move forward explicitly with the 'automated AI R&D workers' framing I introduced earlier. With this in mind, the sceptic in this section is trying to establish the following: although automated workers would be capable of performing AI R&D tasks, it would be *harder* for them to make valuable contributions to AI R&D than the believer has imagined.

If the believer were to adopt something like the 'suite of services' framing instead, she would need to adapt some details of the argument. For example, one might expect the most relevant 'situational constraints' in this story to be those which limit the speed that the ***human supervisors*** work at, rather than those which affect the AI systems themselves (i.e., maybe humans are acting as more of a bottleneck at each step of the recursive process here). Or, if the most relevant constraints are still the ones facing the AI systems themselves, these constraints might just be of a different nature to the ones I describe in this section, since the demands on an AI R&D service or tool are probably different to those on an automated worker.

I won't cover the details of this adapted argument, since my focus is on the 'automated worker' framing. However, I note that this argument would at least be structurally similar to what follows.[79]

The case against short timelines

With the 'automated AI R&D worker' framing of direct recursive improvement in mind, I'm now going to walk through an argument that goes as follows:

- In general, the ease of making capabilities improvements (e.g. via contributions to AI R&D) is not just a function of existing capabilities; instead, these improvements are seriously constrained by situational factors.
- Automated AI R&D workers will therefore require their situations to meet

[79] Tentatively, I think this kind of argument could be slightly more convincing (or at least *more obvious*) in the 'suite of services' case than in the 'automated worker' case. Supposing that human supervisors really are the main bottleneck in this case, because of their role in selecting and synthesising the outputs of AI systems into useful contributions to the field, it's then fairly easy to point to some relevant constraints on those humans that would likely cap the speed of capabilities improvements. For example: they can't retain that much information, it takes them a certain amount of time to learn new information or master new tools, they can't work all hours of the day, they get tired, they get sick, their brain processing speeds are capped, and they can't physically move beyond a certain speed).

But importantly: these are constraints that the AI R&D field, which is currently dominated by humans, *is already facing*. In the 'suite of services' story, humans play *less* of a role in AI R&D than they do currently. The introduction of these AI R&D type direct feedback loops seems to be a big improvement for the field, and progress might still speed up significantly as a result. The impact of this line of argument on the overall debate is thus limited.

However, if we can argue that the AI systems themselves in our chosen DRI story are *also* constrained in some analogous ways to humans (although those constraints are likely to be weaker), we might have some reason to believe that progress actually *won't speed up very much* upon their introduction. This is why I believe that arguments which are based on the situational constraints facing AI systems are much more interesting to this debate.

certain conditions if they are to make significant improvements to AI capabilities.

- Meeting these conditions may be difficult and time consuming.

If this sceptical argument is convincing, it could provide support for (S1) or (S2). That is, it could establish one or both of the following:

- *AI progress from DRI isn't very fast*. Due to constraints or requirements at each improvement 'step', each jump in AI capabilities is considerably smaller or takes much longer than has been imagined.
- *AI progress from DRI cannot be sustained for very long*. At some nearby point in the series of improvement steps, the constraints or requirements faced will be prohibitively strong, such that AI capabilities progress will effectively plateau.

Either way, the believer's case for short timelines via DRI would be undermined.

"In general, the ease of making capabilities improvements is not just a function of existing capabilities levels." Automated AI R&D workers *making improvements to AI capabilities* is an instance of these AI systems not merely possessing a certain ability, but actually *achieving something* within the field of AI R&D.

The sceptic might argue that there are important constraints on what an AI system, given a certain level of capability, *can actually achieve* in such fields. A specific corollary to this would be that there are constraints on the extent to which an automated worker can make improvements to AI capabilities.

The sceptic can make this argument by pointing to other intelligent beings. It seems, in general, that an intelligent being possessing a certain level of (raw) capability will not always display proportionate levels of achievement. For instance, Chollet considers the human case, taking IQ as a rough indicator of raw capabilities in humans. Although correlation between human IQ and achievement has been observed up to a certain point, this relationship seems to break down around higher IQ levels: in particular, human achievements (in scientific fields as well as others) are not overrepresented amongst those with IQs of 170 in comparison to those with IQs in the 120s or 130s.[80]

Chollet notes that, if capabilities did translate proportionately into achievement levels, "then exceptionally high-IQ humans would already be displaying proportionally exceptional levels of personal attainment; they would achieve exceptional levels of control over their environment, and solve major outstanding problems– which they don't in practice".

Chollet argues that this breakdown in correlation is *because our achievements are constrained by our situations*: what a human of a certain level of capability will actually achieve is dictated by the physical environment, historical and social context, and so on, in which she exists. Without the right conditions, she will not succeed in making significant contributions in her field.

[80] Chollet doesn't back this point up with studies, but illustrates it by way of some examples: "many of the most impactful scientists tend to have had IQs in the 120s or 130s — Feynman reported 126, James Watson, co-discoverer of DNA, 124 — which is exactly the same range as legions of mediocre scientists. At the same time, of the roughly 50,000 humans alive today who have astounding IQs of 170 or higher, how many will solve any problem a tenth as significant as Professor Watson?"

Importantly, even if one doubts this point, the rest of the proceeding argument (which is about the senses in which automated AI R&D workers might be constrained by their situations) might still be accepted; the IQ point is primarily intended to serve as a helpful analogy.

(To clarify: capabilities are certainly an input to achievement, but above a certain level of capability in humans, it seems they are no longer the *primary bottleneck* for achievement. Instead, our achievements become bottlenecked on features of our situation.)

"Like humans, AIs require certain conditions to make improvements to AI capabilities." Chollet believes that the above reflections are not just applicable to humans, but are generally true for 'intelligent' beings, including AIs. That is, to make significant contributions to any field, AI systems (like humans) must have the right conditions. This applies to the field of AI capabilities progress, and means that the specific process of automated workers *making significant contributions to AI R&D* could also become bottlenecked on certain features of their situations.

This claim might seem odd at first. It's certainly true that *humans* need a variety of external factors working in their favour in order to achieve just about anything. But many of these factors seem plainly irrelevant to the AI context: for example, there are no clear AI analogues to the socio-economic dynamics, cultural practices, physical or mental health conditions, or personal life events that shape a human's achievements in practice.

These factors, which seem unique to the human context, are not what the sceptic has in mind. It appears that there are other, more universally applicable features of human lives that have empowered some of us to contribute to research, develop new technologies, and so on, and which may be broadly relevant for enabling similar achievements in all 'intelligent' beings.

Drawing on some key themes from Chollet's work, these features may perhaps include: interaction and coordination with other intelligences; a long history of collective efforts; access to and understanding of a vast body of knowledge; strong familiarity of the individual with her specific environment; and the ability to use tools within it. An automated worker of some given capability level might plausibly need similar conditions to these in order to successfully contribute to AI R&D.

What are the relevant conditions for automated worker 'success' in AI R&D? There's no definitive list of these conditions from Chollet's article, but I break down the kinds of requirements his arguments *might be taken to imply* under three categories:[81]

- Requirement A: The automated worker interacts and coordinates work with a large number of other intelligences.
- Requirement B: The automated worker grasps a large amount of relevant information, including a body of knowledge, previous work in the field, and historical context.
- Requirement C: The automated worker engages, and becomes familiar, with its environment (including having access to and an ability to use relevant tools, and possibly having sensorimotor affordances).

A sceptical argument can now be made with direct reference to these

[81] Some examples (but not all!) of the support I find in Chollet's work for these particular requirements include the following remarks from *The implausibility of intelligence explosion*:

- "[inventions require] billions of brains, accumulating knowledge and developing external intelligent processes over thousands of years, implement[ing] a system — civilization"
- "your sensorimotor affordances ... are a fundamental part of your mind. Your environment is a fundamental part of your mind. Human culture is a fundamental part of your mind". (Here, the phrase "your mind" is being used roughly to capture *the thing that makes you 'intelligent' and thereby enables you to achieve things, solve problems, and so on*.)

requirements. The sceptic could say something along these lines:

> There's no guarantee that an automated AI R&D worker with a given level of capability would be immediately successful at making further improvements to AI capabilities. Indeed, doing so would require meeting requirements such as A, B, and C. It's not that these requirements are in principle impossible to meet, but that creating these conditions for making significant capabilities improvements at each step of the recursive process could be time consuming and difficult, or even become resource-prohibitive. As a result, improvement steps could be smaller or slower than the believer has imagined, or could hit a plateau.

How hard is it to meet these requirements? I'll go through each proposed requirement in turn.

Requirement A: Interacting and coordinating with other intelligences

This seems easily achieved at each step of the recursive process. We've already seen examples of AIs collaborating with one another. And, given the number of copies of an AI system we could plausibly make (as well as the possibility of continuing to have some humans working on AI R&D), it does not seem hard to give automated workers an opportunity to interact and coordinate with other intelligences in the process of making AI capabilities improvements.

In fact, this set up is already part of my chosen story of direct recursive improvement: proponents of this story typically imagine large numbers of automated workers, not working in isolation from one another, but in a coordinated way.

Requirement B: Grasping lots of information

This might also not be very difficult to achieve at each step of the recursive process. Current AI systems based on neural networks with deep learning are already able to digest (and make use of) huge amounts of information very quickly, due to their fast processing speeds and ability to run non-stop for long time periods. For example, we now have LLMs such as ChatGPT which have essentially been trained on the entire internet. (It should be noted that Chollet wrote the referenced article in 2017, before this had been achieved.) As another key example, AlphaGo Zero learnt to play Go at a beyond-expert level within a few days.

The successes of these AI systems also indicate a difference in the training needs of AIs compared to humans: it is not always necessary to confer information to AI systems *in a similar way to humans* to achieve the same (or better) performance. For example, AlphaGo Zero was essentially self-taught. It was able to bypass centuries of historical Go knowledge and context, relying instead on learning through self-play. On this point, Yudkowsky notes that:

> "If AlphaGo Zero can waltz past all human knowledge of Go, I don't see a strong reason why AGI Zero can't waltz past the human grasp of how to reason well, or how to perform scientific investigations, or how to learn from the data in online papers and databases."

Of course, this doesn't mean that each generation of new automated workers under a DRI scenario won't need lots of information. Indeed, the first version of AlphaGo needed to be initialised on a vast dataset of played games, and AlphaGo Zero had to play 4.9 million games against itself (in rapid succession) to surpass expert level. However, it seems that AIs aren't restricted to the same route of acquiring knowledge as humans typically take. As such, some of the most time-consuming and complex features of the situation required for human learning – e.g. the requirements that learning occurs gradually, in a certain order, and often over centuries of human history (or at least years of a single human lifetime) – may not really apply in the AI case. AI systems find shortcuts.

Requirement C: Engagement and familiarity with the environment

By this, a sceptic might mean the ability of an AI to interact with and effectively shape its environment, and adapt its responses to it.

In the context of an automated worker making contributions to AI R&D, this might break down into a few possible requirements.

(i) *Automated workers need access to relevant tools, and an ability to use them.*

This appears to be a very plausible requirement. To make capabilities improvements to AI systems, humans require certain tools. It seems likely that an automated AI R&D worker would need to use similar tools to a human working in the AI R&D field.

The most plausible contenders for tools that would be helpful in this context are the internet, search engines, books and papers (as digital files), code libraries, and various software tools.

Enabling automated workers to access such things and effectively use them does not seem like a difficult barrier. In fact, current AI systems are already demonstrating promising use of many of these tools: ChatGPT can now interact with plugin apps (e.g. it can query Wolfram and Midjourney), use search engines, and read uploaded files to produce responses to prompts; more recently released systems Devin (dubbed "the first AI software engineer") and Replit Agent are both able to access code libraries to build apps.

(ii) *Automated workers may need sensorimotor affordances (i.e. capacity for both sensory and motor activities).*

This would involve some form of physical or virtual embodiment, through which automated workers could receive sensory inputs (or equivalent) and react to them, and physically interact with any features of the environment that are relevant to its tasks.

This is a major part of how humans learn, solve problems, and achieve things. However, its relevance on the pathway for developing advanced forms of AI is

very speculative, and has been contested. (See my commentary under 'embodiment as a bottleneck' in the previous chapter, as well as wider debates about embodied cognition).

It seems *especially* unconvincing as a proposed requirement in the specific case of contributing to AI R&D. Interacting with a physical environment does not *seem* to be essential for success in this context (barring certain manufacturing needs, which could continue to be met by human workers in the near-term while robotics catches up): given access to certain tools, most of the activities involved in AI R&D can probably be done in a remote, sensorless, disembodied way.

To whatever extent sensorimotor facilities or embodiment actually becomes necessary for automated workers to sustain the process of DRI, meeting these requirements would not present an *insurmountable* challenge: with efforts into robotics or into creating detailed simulations, it seems that they would be possible to address. Of course, this could still delay AI progress, and thereby reduce the likelihood of a short timeline.

(iii) *An automated worker may need time to become familiar with its environment and develop a mastery of the affordances it has within it.*

At each step of the process of recursive improvement, automated workers making further improvements might not just be a case of *having* the tools and faculties described above, but becoming *sufficiently skilled* with those affordances to successfully manipulate their environment according to their goals.[82] It's unclear what meeting this requirement would look like, in practice. Speculatively: it might require these automated workers to take repeated actions and receive feedback from their environment[83] (perhaps in the form of conducting experiments, testing predictions, and so on); it might also involve each automated worker possessing some model of the relationship between its environment, its affordances, and the task at hand. Whatever this looks like exactly, it might take time to achieve.

In this vein, Chollet compares the development of advanced forms of AI to the act of putting a human brain in an octopus body, asking:

> "What would happen if we were to put a freshly-created human brain in the body of an octopus, and let in [sic] live at the bottom of the ocean? Would it even learn to use its eight-legged body? Would it even survive past a few days?"

The relevant point here is that a human brain, given an octopus body and other octopus affordances, wouldn't easily or immediately be able to solve tasks in an octopus environment. (In fact, Chollet questions whether this brain would be able to solve such tasks *at all*.) Likewise, an automated worker, given certain tools and affordances to utilise in a new environment, might still struggle to master those tools and affordances to the extent necessary to succeed in this environment.

Let's assume it is *in principle* possible for an AI system to succeed, in the

[82] Although this idea fits most naturally as a subpoint of requirement C, we might also view it as a subpoint of requirements A and B: that is, to make further capabilities improvements, an AI system might also need to become familiar with, and develop a mastery of, its opportunities for interaction and the information it can access (rather than just possessing those opportunities and that information).

[83] This calls to mind Hubert Dreyfus' notion of 'skilful coping', and his work on the importance of embodied-embedded cognition in the AI context.

relevant environments, at the task of making AI capabilities improvements – and the bottleneck for this is not having the appropriate tools and affordances, but developing sufficient mastery of them. It's still not clear what, exactly, achieving this mastery would consist in, but the sceptic of short timelines might argue that it would at least be time consuming to achieve. It might also be necessary for automated workers to develop new familiarity and mastery *at each step* of the recursive improvement process: after all, the set of challenges facing these AIs, tools available to address them with, and so on, will likely change over time. As a result, we might expect this to translate into repeated delays in making capabilities improvement steps.

The believer can respond to this argument fairly easily. Firstly, she can point out that the octopus analogy is misleading. The process of recursive improvement in AI development will not involve just dropping some disembodied AI brain into a random new environment, with a random new body or set of tools, and asking it to learn what to do. Instead, at each step of the process, automated workers can be designed and trained *for* their specific situation, taking into consideration the tasks they will face, the tools they will have on offer, and any physical bodies they will be given (at least to the extent such things can be pre-empted, and relevant training data can be supplied).

Moreover, the successes of LLMs like ChatGPT, Claude, and Gemini illustrate that many AI systems are already successfully mastering the situations that they are specifically designed for, without being given additional time beyond their initial training period to develop this mastery. (They might *benefit* from some kind of additional learning time, but it seems they haven't *needed it* to achieve many impressive feats.) These LLMs are trained on huge datasets which closely capture the situations they will be faced with once deployed, such that they can then immediately start succeeding at their tasks in the deployment environment. The only 'further learning' that many existing LLMs have even been given the capacity for is via a context window (which is just a temporary store of information),[84] yet these systems can achieve impressive performance on a wide range of benchmarks and in a wide range of contexts.

It's possible that *future* AI systems (tasked with performing more complex actions, some of which may be involved in AI R&D) will need to develop further 'mastery' after training, even if that hasn't been a necessity for previous systems. But to the extent that this would be necessary, it might not involve *significant* additional time requirements or any major breakthroughs in system design that would seriously delay the arrival of TAI. For example, dialling up run-time compute in current AI systems might help to unlock meaningful new levels of skill after training, in a sense that might be relevant for situational 'mastery'. Alternatively, there could be some relevant improvements from just increasing a model's context window length. Such alterations could slow down the pace of LLM output production (or input processing) to some degree, but it's not clear how this effect could be significant enough to result in these systems taking, say, additional *months or years* to make capabilities improvements at each step of the recursive improvement process. And even if

[84] Some recent AI models featuring base LLMs, like OpenAI's o-series models, are now enjoying something that could be seen as 'further learning' after deployment, in the form of increased run-time compute. Like the information 'learned' through a context window, this is also temporary: since the model does not update its weights during the run-time period, any understanding developed through e.g. chain of thought reasoning must be relearned from scratch by the model for each new problem it is faced with.

The important point here is: even in these groundbreaking cases of success, 'mastery' in the sense of *permanently* updated knowledge of an environment or problem still hasn't been required. Perhaps this is something we would still label as a form of mastery, but it seems less mysterious and time-consuming to confer than a sceptic might suggest, as I indicate in the paragraph below.

this did come with serious time requirements, it would be as a tradeoff for the improved *quality* of LLM outputs, so the capabilities improvements made at each step of the process might be much *greater*. As such, it seems that the overall impact on the timeline to TAI would probably be minimal.

There are two key takeaways from this: firstly, (as of late 2024) there's little evidence for additional learning being *essential* for an LLM to 'master' its given situation, beyond the training time requirements already incurred under requirement B (i.e. those associated with data). Secondly, if there is any need in the future for automated workers to develop some further 'mastery' after training, it's not obvious that this would cause any *significant* delays to the process of recursive improvement.

Taking stock, and noting a possible reframing

The sceptical argument, as advanced above, has not been hugely compelling. Even if the sceptic is largely right that requirements A, B, and C are necessary for automated workers to succeed in making further AI capabilities improvements, it seems that meeting those requirements might not actually be very challenging or time-consuming to meet. (The exception to this might be the sensorimotor aspect of C, but this is the most speculative of the suggested requirements.)

In particular, the arguments above have not stood up well to evidence from recent progress in AI capabilities. Over the past few years, LLMs have enjoyed notable successes in solving the specific tasks that face them – and in doing so, they have shown an ability to navigate their environments, utilise tools and other affordances available to them, learn the relevant context, interact with others where necessary, and so on. In light of this, it's hard to see why meeting these requirements would be particularly difficult in the future (or, to the extent that any requirements *are* difficult to achieve, why meeting them would be *necessary* for LLMs to be successful at their given tasks, if they haven't been necessary so far.)

It is not surprising that these arguments appear shaky in light of recent developments in the field, since my framing has drawn primarily from a 2017 article from Chollet (which pre-dates, for example, the release of GPT-2) and has also been silent on any details related to the paradigm to which the posited systems belong.

This might motivate the sceptic to adapt her argument to bear more directly on the *specific limitations of LLMs*. For example, I think she could frame the difficulty of automated workers making capabilities improvement steps in terms of the difficulties traditional LLMs appear to have with generalising to meaningfully new problems and situations, or with learning in a dynamic way. These are both limitations of traditional LLMs that Chollet himself has emphasised in more recent discourse (for example, in his podcast appearance with Dwarkesh Patel).

Such an approach would involve a continuation of certain arguments

previously brought out in the compute scaling chapter of this report: that a paradigm shift is needed before TAI can be developed, due to fundamental limitations of traditional LLMs. In that chapter, the sceptic argued that traditional LLMs will not easily reach TAI via compute scaling; now, one might try to extend or adapt similar considerations to argue that they won't quickly reach TAI via recursive improvement, either. I will not pursue this line of reasoning here, given the diminishing relevance of critiques of 'traditional' LLMs in the wake of OpenAI's (somewhat untraditional, and very promising) o-series models.

Capabilities improvements get more difficult at each step

The previous objection said: "each step is hard". This objection says: "each step is harder than the last".

The basic objection

The argument for short timelines via fast DRI goes like this: direct feedback loops in AI R&D mediate step-by-step capabilities improvements. At each step, the newly designed 'R&D AIs' (defined in Figure 2.2) *get better at driving AI capabilities improvements* than their predecessors. This could result in the acceleration of actual improvements in capabilities at each step.

The sceptic might agree that R&D AIs become (in some sense) better placed to make or support capabilities improvements at each step of the recursion – but point out that, at the same time, *capabilities improvements are getting increasingly difficult to make*, as the growing AI R&D field faces certain bottlenecks and counter-reactions.

This phenomenon might slow down the pace of actual AI capabilities progress, or even present such an obstacle to subsequent recursive improvement steps that capabilities progress plateaus completely. Progress may therefore, in practice, not be fast or sustained enough to produce TAI within the next ten years.

In the words of François Chollet: "Exponential growth, meet exponential friction!"

A note on this objection. In the previous section, I outlined reasons to think that progress in AI R&D could be constrained at each step of the recursion. This might mean there are *slower capabilities improvements overall* than the short timelines believer has argued for. In this section, the sceptic is making a slightly different claim: that constraints increase at each step of the recursion, such that *capabilities improvements get increasingly difficult*. That is, there are certain decelerating factors which essentially cancel out some of the accelerating effects of positive feedback loops.

The difference might be put in this way: the sceptic from the previous subsection said "each improvement step is hard", while this sceptic says "each improvement step is harder to make than previous ones".[85]

[85] Of course, both sceptics might sometimes appeal to the same constraints on capabilities improvement when making these arguments. Indeed, the final two bullet points in the list of 'decelerators' below are related to the sceptic's requirements A and B from the previous subsection.

Possible 'decelerators' of recursive improvement

Below, I list several reasons to think AI capabilities progress might get harder at each improvement step. The presence of these decelerators could mean that the progress corresponding to recursive improvement is slower than imagined, or that the trajectory of recursive improvements eventually reaches a plateau.

It should also be noted that many accounts of short TAI timelines via recursive improvement do already take into consideration, to some degree, the effects of some of the below decelerators; I'll revisit this point in 'Who wins the tug of war?'.

Note: Although the 'automated workers' framing of DRI remains my focus for this section, the decelerators below seem generally applicable to scenarios featuring any 'R&D AIs' (except where specified otherwise).

- **New insights get harder to come by** as low-hanging fruit in capabilities improvements is quickly exhausted. As a result, it may take increasing amounts of effort to even sustain current levels of progress, or perhaps the recursive improvement period simply fizzles out as we run out of ideas.
- **We face diminishing returns from algorithmic research** as we approach the best achievable performance on certain algorithms. For example, as David Thorstad points out in *Against the singularity hypothesis*, "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" and "if even one process resists being sped up, the whole system of recursive self-improvement may be delayed."[86]
- **Physical demands of AI development get harder to meet at each recursive step.** Rapidly increasing sizes or numbers of R&D AIs will translate into increasingly high demands on physical compute, materials, energy, and so on, as well as the supply chains through which labs can access these things. I already noted these as bottlenecks to AI progress in the chapter on compute scaling.
- **Coordination between researchers gets harder** as the field of AI R&D gets much bigger. *Note that this makes sense under the 'automated workers' framing of DRI, in which large numbers of AI systems are deployed to make capabilities improvements, and the number of workers that are active in the space may be increasing with each new generation. It may present less of a challenge in DRI scenarios featuring a 'suite of services'.*
- **Training and education becomes more difficult and time consuming** as the body of relevant knowledge in the field (the literature, past experiments and models, and so on) expands. Each generation of R&D AIs has to get up to speed with some amount of this knowledge in order to make further capabilities improvements; this could get increasingly difficult.

[86] It is, of course, contestable that we are nearing best achievable performance on search algorithms. It's also not clear that significant progress on search algorithms will actually be essential for future capabilities improvement, anyway. Indeed, many experts expect improvements on learning algorithms to be the driving force behind algorithmic efficiency in future. There may be plenty of room for improvement on this front: Aschenbrenner, for example, suspects that there are currently at least 5OOMs of algorithmic improvements still available to us. (Still, the sceptic might think we'll exhaust those OOMs before TAI is achieved.)

A possible response to the 'steps get harder' objection

In response to some of the above challenges, one might argue that at each step of recursive improvement, automated workers will not just be getting better at making AI capabilities improvements: they will likely also be becoming more *generally* intelligent (in the sense relevant here, more capable of responding to a *wider* range of challenges[87]). This could empower them to break bottlenecks.[88]

In this vein, Yudkowsky has argued that "smart agents will try to deliberately bypass these bottlenecks and often succeed. This is why the world economy continues to grow at an exponential pace."

Earlier in this chapter, I speculated that the effects of positive feedback loops in AI development might help us overcome some of the bottlenecks described above, such as those relating to physical compute and algorithmic progress. This was not to say that such bottlenecks *won't be a problem for recursive improvement at all*, or won't affect the likelihood of a short timeline. It's worth briefly describing one way this 'bottleneck-breaking' process might go:

- At some stage of the recursive improvement period, challenge X from the above list is bottlenecking AI capabilities progress.
- The effect of challenge X is that each capabilities improvement step is slowed down, perhaps even by greater amounts at each step.
- However, improvements do continue, albeit slowly.
- After some improvements, R&D AIs reach a new level of capability or general intelligence at which they are now able to address challenge X.
- Challenge X thus has less of an effect in future capabilities improvement steps.

This would certainly amount to a delay in AI progress, but a temporary one – perhaps not quite the sustained "exponential friction" that Chollet has warned us about.

Issues with the 'bottleneck-breaking' response

It's tempting to treat 'bottleneck-breaking capability' as some silver bullet, but its applications may in fact be limited.

Thorstad makes a useful distinction that I will appeal to here. Suppose an agent is tasked with solving some problem. Some of the challenges of achieving this are *features of the problem*, while others are *features of the agent* herself. (Compare: "the problem has x, y, z areas of complexity" vs "the agent lacks skills a, b, c".)

The response above, from the short timeline anticipator, effectively tells us that there is something special about the agents[89] in the AI case: they can be made increasingly better at problem-solving.

But some of the sources of friction in the list above don't *just* come down to features of the agent; they are, at least in large part, underpinned by some

[87] Generalising to *unseen* challenges, as in the definition of 'generality' I provided in the chapter on compute scaling, may not be necessary for this particular argument to hold.

[88] This at least seems true under the 'automated workers' framing of DRI, which is my chosen framing. Something similar might be said in the 'suite of services' case: we might imagine the composite system of humans + AI services becoming overall better placed to respond to a growing range of challenges, even if each individual service remains fairly narrow in the scope of its tasks.

[89] Under my preferred framing, these agents are just the automated AI R&D workers. Under the alternative 'suite of services' framing, they are perhaps best seen as the composite system of many-AI-services-plus-human-supervisors.

feature of the problem. For example, Thorstad argues that the exhaustion of low-hanging fruit is primarily an issue about the solution space.[90] Others on the list, such as the prospect of approaching best-achievable performance and some of the increasing physical demands, might also be viewed similarly: they involve some inherent difficulties of the solution space such as the non-existence of superior algorithms or the physical limitations of hardware.

In these cases, the sceptic might claim, making the agent better at problem-solving is only going to get us so far towards actually solving the problem; it might help to an extent, but it ultimately won't change the underlying features of the solution space that are challenging to address. And even if this is only true for one of the bottlenecks highlighted above, this could still count against the likelihood of fast recursive improvement. As Thorstad notes:

> "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."

The relationship between accessible input improvements and capabilities improvements is sublinear

Here, it will be helpful to distinguish *the inputs to AI capabilities improvements* from *the AI capabilities improvements themselves*.

Those who expect short timelines via recursive improvement sometimes argue that feedback loops will produce fast improvements on certain inputs to capabilities development. Some of the most likely inputs to be rapidly improved in DRI scenarios are processing speed, memory, search depth, and (effective or physical) compute.

Even if this is true, however, it might not point towards a short timeline. David Thorstad has argued that, in general, AI capabilities do not improve as quickly as these inputs to capabilities progress. In other words, he claims that capabilities *grow sublinearly* in relation to more 'accessible' input improvements. As such, fast improvements on these inputs won't necessarily yield fast improvements in actual capabilities.

To illustrate this visually, Thorstad's claim is that the relationship between capabilities and accessible inputs in AI development looks something like this:

[90] In fact, it seems to be a feature of solution spaces in general, rather than something unique to the AI field. As Thorstad notes, (italicisation mine): "many social scientists think that beyond a point, *nearly all idea-generating processes* behave like fishing" in the sense that "after a while, any sensible investigatory process will have made more than its share of the easiest discoveries, and subsequent discoveries will become harder".

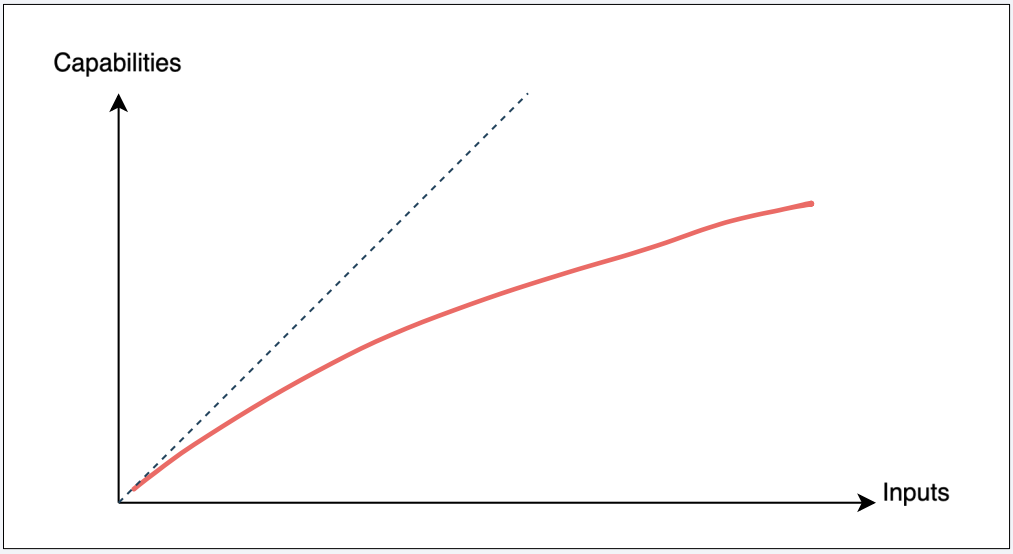


Figure 2.7: A sublinear relationship between accessible inputs and capabilities

As a result, even if the trajectory of input improvements is fast, capabilities improvements may not be. For example, we might expect to see something like this:

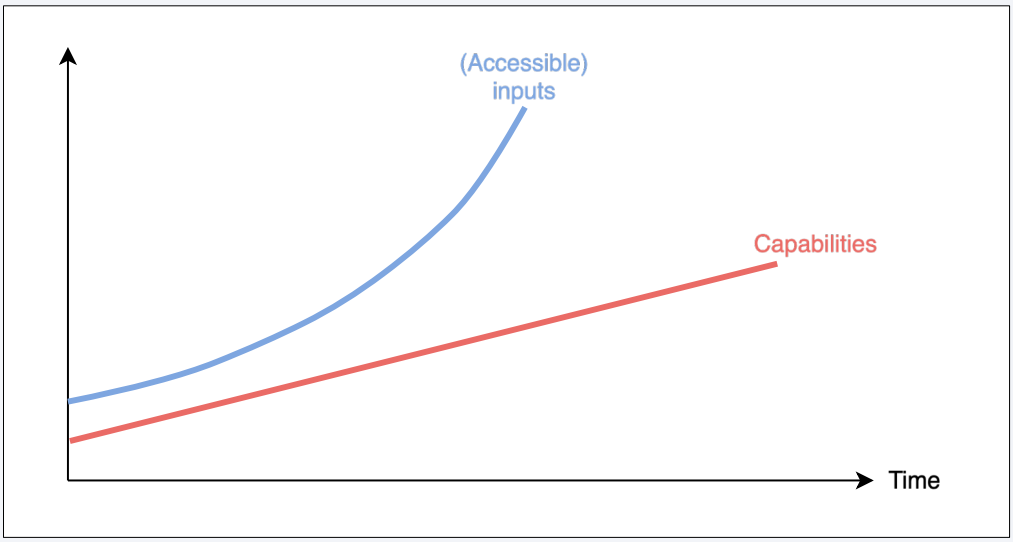


Figure 2.8: Comparison of the (possible) trajectories of accessible input improvements and capabilities improvements over time.

To defend this claim empirically, Thorstad appeals to several examples, including the apparently sublinear relationship between capabilities and compute in the development of Chess- and Go-playing systems. Citing a 2022 study from Thompson et al., he comments:

> "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."

If this observation is indicative of a wider phenomenon, and we can generally expect capabilities to grow sublinearly in relation to compute and other accessible inputs, what are the implications for the arguments at hand?

The apparent upshot is that the rapid, 'explosive' progress in AI capabilities that has been envisioned by many believers in short timelines via recursive improvement could actually be harder to achieve than they have imagined. Importantly, *however fast* we believe the trajectory of AI capabilities progress would need to be to take us to TAI by 2035, it might need to be driven by *even faster progress* on certain relevant inputs. For example, to achieve exponential growth in capabilities, Thorstad believes you will need *superexponential* growth on some driving inputs. "And [to achieve] hyperbolic growth [in capabilities]? Don't ask."

Possible responses to the 'sublinearity' objection

One way of responding to this would be to object on empirical grounds. The sceptic could point to a period of time in which exponential growth in certain inputs actually *did* correspond to exponential growth in capabilities, or she could supply a point-in-time example of a large, discontinuous jump in capabilities (that didn't correspond to a similar jump in inputs). Neither example would completely falsify Thorstad's claim of sublinearity: perhaps if we zoomed out to view the relationship between capabilities and inputs on a large enough timescale, those examples could just amount to noise on a graph that is overall sublinear. However, in the debate over short timelines, we're interested in a scale of ten years or less. Within this short time frame, any period of exponential growth or discontinuous jump in capabilities could be noteworthy: if fast enough or big enough, it could potentially cross the remaining distance to TAI. Therefore, any counterexample illustrating that such an event is possible could be a powerful response in this context.

I won't try to supply such a counterexample. Instead, let's suppose Thorstad's claim has been more or less verified empirically. I want to briefly question *why* this sublinear relationship holds, to understand if it could be changed or overcome through the efforts of AI R&D.

As evident from the graphs above (Figures 2.7 and 2.8), Thorstad's claim of 'sublinearity' effectively amounts to a claim that there are diminishing returns to capabilities improvements from accessible input improvements.[91]

I think the most obvious explanation for this echoes several arguments from the compute scaling chapter and earlier counterarguments in this chapter: that capabilities improvements are bottlenecked on factors which are simply harder to improve than these more 'accessible' inputs, and which are likely to become increasingly difficult to improve over time. As such, improving the more accessible inputs does not straightforwardly or continually result in proportionate increases in capabilities.

If that is the conceptual basis, then the argument in this section *might just collapse into arguments I made earlier on*. The sceptic could perhaps respond,

[91] Note that is *slightly* distinct to the arguments under 'Capabilities improvements get harder at each step' that there are diminishing returns to capabilities improvements *from effort*, due to basically exhausting low hanging fruit in input improvements or approaching best achievable performance on certain inputs. These challenges appear to be somewhat more fundamental to the solution space, as noted, while the challenge at hand (diminishing returns to capabilities improvements *from input improvements*) might be more easily overcome.

as she did before, with some suggestions of ways in which the AI R&D field could get unblocked on the relevant challenges. Perhaps specific bottlenecks such as data can be overcome; and perhaps each new generation of R&D AIs deployed to improve AI capabilities, given their own increased generality or problem-solving abilities, would be able to increase the overall input-efficiency of capabilities improvements. This could help to bridge the gap between inputs and capabilities and thereby counteract diminishing returns.[92]

Progress in narrow vs broad domains: reflections from the field of economics. The ideas of sublinearity and diminishing returns in AI capabilities progress (ideas which are central to this counterargument and to previous ones) connect strongly with existing research in the field of economics, such as Bloom et al.'s *Are ideas getting harder to find?*. From this body of literature, it appears that progress with respect to any specific benchmark within a narrow domain will eventually be saturated, resulting in diminishing returns. This applies to Thorstad's selected examples of "Chess and Go ELO rating, ability to solve protein folding problems, or to predict weather patterns and oil reserves", and is consistent with the sublinear trends he identifies there.

However, in the debate over TAI timelines, we are not solely interested in capabilities progress towards specific narrow targets; since something like *general intelligence* may be key to unlocking transformative capabilities, we are also interested in progress more broadly, including the development of new capabilities and the expansion of AI applications into new domains. Thorstad may be correct that with consistent improvements on accessible inputs, progress towards certain benchmarks has plateaued – but at the same time, AI capabilities have expanded into a variety of new fields, with individual systems becoming increasingly general in their applications. Thorstad's observations therefore don't seem to seriously dampen the prospect of *broader capabilities progress*, still based on improvements in accessible inputs via DRI, taking us quickly towards something like AGI.

Further reading (on all three counterarguments):

- *François Chollet - The implausibility of intelligence explosion*
- *François Chollet - On the measure of intelligence*
- *Eliezer Yudkowsky - A reply to François Chollet on intelligence explosion*
- *David Thorstad - Against the singularity hypothesis*
- *Chapter II of Leopold Aschenbrenner's Situational Awareness*

[92] Perhaps I have missed some other, more fundamental reason for sublinearity that does not reduce to some similar case of bottlenecking, and which could not in principle be addressed in similar ways to previous counterarguments.

Who wins the tug of war?

These three counterarguments are not decisive against short timelines

It's important to acknowledge where the debate actually lies here. It's not the case that these counterarguments, if valid, are decisive against short timelines. Indeed, those who expect short timelines via recursive improvement can, and sometimes do, factor in many of the above challenges into their models.

What really differentiates the members of this group from the sceptics here is that even if they accept that these challenges will affect recursive improvement dynamics, they just don't believe that this will slow progress down *that much*, or mean that capabilities improvements will plateau any time soon.

On the 'each step is hard' objection: The believer could accept that there are constraints on the achievements of automated workers at each step of the recursive process, such that certain conditions must be met if automated workers are to stand a chance at making further contributions to the field. However, she will simply argue that these conditions are not actually very onerous to meet, and might therefore not slow down or reduce the size of each capabilities improvement step by a significant amount.

On the 'steps get harder' objection: The believer could accept the sceptic's arguments that capabilities improvements will get fundamentally harder with each step, but maintain that the ability of R&D AIs to improve capabilities will just be *increasing way too fast* for this to really matter. Indeed, Aschenbrenner accounts for this in his argument for an intelligence explosion. With regards to the claim that new insights in capabilities research will get harder to come by, such that increasing effort is required to sustain progress, he states that:

> "I think this basic model is correct, but the empirics don't add up: the magnitude of the increase in research effort—a million-fold—is way, way larger than the historical trends of the growth in research effort that's been necessary to sustain progress"

The Epoch team has also incorporated a variable capturing "diminishing returns to discovering new ideas over time" in an analysis of the returns from automating AI R&D designed to better understand the prospect of a software 'singularity'. Even with this variable estimated and accounted for, their empirical findings still "suggest that the returns to software R&D might be high enough to drive hyperbolic growth in software alone, although the evidence is not conclusive".

In a similar vein, Bostrom explicitly factors *recalcitrance* into his model of intelligence explosion, as a measure of the difficulty of making capabilities improvements. He argues that progress could be fast even if recalcitrance is increasing over time, provided that *optimisation power* (i.e. the quality-adjusted research effort applied) grows sufficiently rapidly.

On the 'sublinearity' objection: The believer could accept the sublinearity of capabilities growth with respect to accessible input improvements, but argue that (i) this relationship is only marginally sublinear or (ii) the relevant inputs will be improving so fast that we'll still see fast enough capabilities progress to produce TAI within ten years.

What this means for the debate

From the above reflections, it appears that the main tension over what TAI timeline results from a period of recursive improvement is not about whether these challenges will have *any* slowing or constraining effects on progress, but about *how severe* those effects actually are. *To what extent* will they slow progress? Will they become *strong enough* to eventually cause capabilities improvements to plateau? And *how* quickly (if at all) can they be overcome?

What we end up with, in relation to these questions, is a back-and-forth argument over a variety of posited drivers and posited restraints to AI capabilities progress, pulling us in two quite different directions. The believer will say that the drivers at play in recursive improvement will have, over the next decade, much stronger effects than the restraints, such that the resulting trajectory of capabilities growth will still be fast enough to produce TAI; the sceptic will disagree.[93]

Who wins? To establish this, dedicated efforts to actually quantify the impact of each relevant driver and restraint would be helpful. One recent and exciting effort in the direction of quantifying restraints is Epoch's *Can AI scaling continue through 2030?*, which features estimates of the impact of certain bottlenecks to scaling, including data scarcity. I would be eager to see estimates of this kind specifically incorporated into models of direct recursive improvement, to illustrate how relevant restraints may come to bear on the trajectory of capabilities progress in these scenarios.

It should be noted that Tom Davidson's compute-centric model of takeoff speeds, which factors in the impact of AI R&D automation, also incorporates the effects of several of the restraints to progress via AI R&D automation discussed in this chapter – and yet still places significant weight on short timelines. (See Table 1.1 and the subsection of this chapter entitled 'Will capabilities progress accelerate?' for more on Davidson's work.)

I won't provide a summary here of any other literature that seeks to do something like this. Instead, for the purposes of this chapter, I'll now consider what the trajectory of capabilities improvements *would actually need to look like* for the short timeline believer to win this tug of war. I'll argue here that the plausibility of 'explosive' trajectories isn't really that important to the present debate.

How do you win the tug of war?

Given the complex interplay of drivers and restraints described above, it is reasonable to have at least *some doubt* over stories of AI capabilities progress

[93] Of course, these reflections are not unique to the debate over recursive improvement scenarios, but in fact apply to almost any debate between those who expect short timelines and those who do not. Indeed, I could have made similar comments in the compute scaling chapter of this report. However, it's especially helpful to note this dialectical situation when evaluating the counterarguments of this chapter, which have specifically called into attention the *shape* of the capabilities improvements trajectory that could result from recursive improvement.

which feature explosive, runaway trajectories of growth (such as the fast exponential or even hyperbolic growth modes described by the likes of Bostrom and Chalmers). Taking the sceptical arguments seriously, the capabilities growth associated with recursive improvement dynamics could plausibly just be linear (or quickly become linear, due to growing bottlenecks). Even granting this much to the sceptic, however, would not mean that she automatically wins the debate.

Recall that, at this start of this chapter, I noted that:

> […] if each capabilities jump under recursive improvement is at least *greater* in size than the equivalent jump that human R&D workers alone would have produced in the same time, then we'll see faster growth in capabilities than we would have otherwise. This may not characterise a sustained period of acceleration, but *could* still correspond to a sudden – and possibly dramatic – step change in growth rate.

That is to say: even if AI progress is too strongly constrained to follow an explosive trajectory once a period of direct recursive improvement begins, *it could still be very fast*. For example, the initial emergence of DRI dynamics might bring about a significant one-time speed up in AI progress, resulting in a trajectory like the blue trendline (d) below.[94]

Five examples of possible trajectories after DRI begins

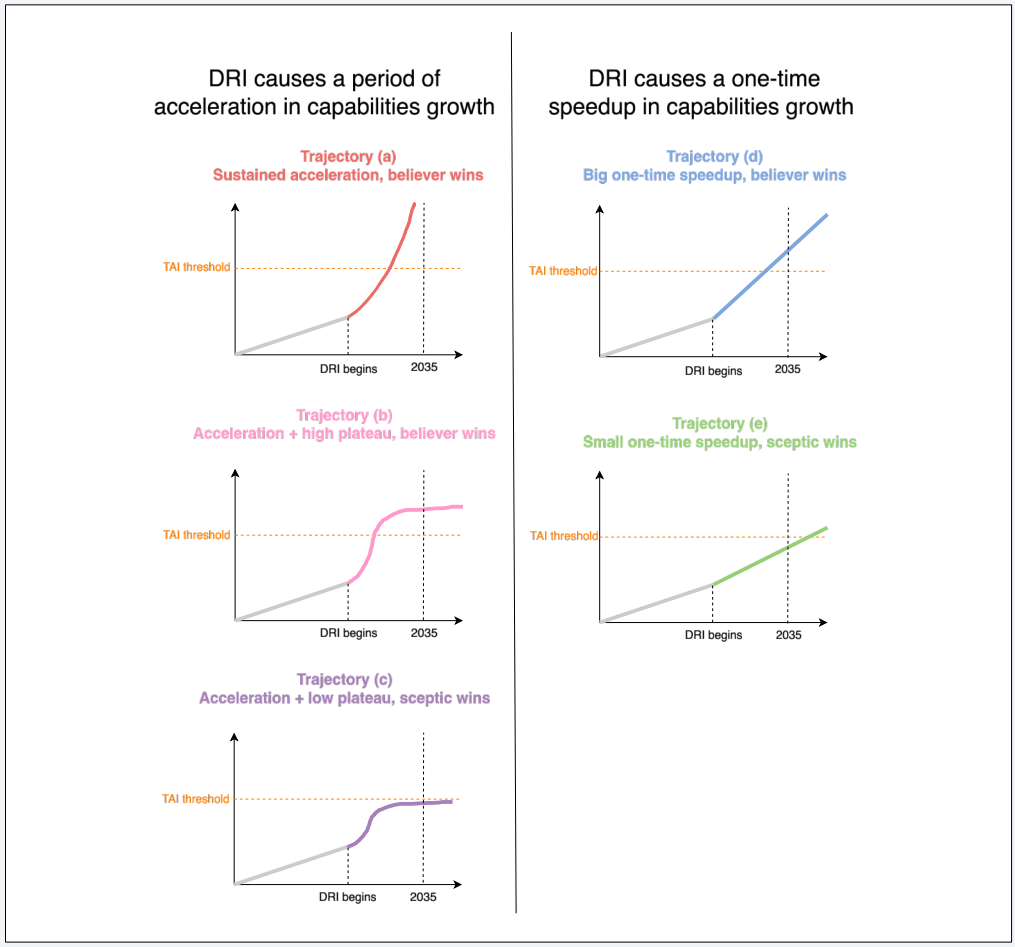


Figure 2.9: Five examples of possible trajectories after DRI begins. The graphs above assume that some form of DRI begins at some fixed date before 2035, and illustrate five different options for

[94] In fact, even a step change in growth rate might not be required. Recursive improvement dynamics being strong enough to simply sustain current rates of capabilities progress (say, by breaking bottlenecks to compute scaling) might indeed be enough to bring about TAI within the next decade. To see this, imagine that the threshold for TAI is slightly lower than that shown in Figure 2.9 or that DRI begins slightly sooner.

I'll return to this idea in Chapter 3, when I come to generate 'short TAI timeline scenarios'.

the subsequent trajectories of AI progress. In examples (a), (b), and (d), the timeline to TAI is short (the 'believer' wins); while in (c) and (e), it is not (the 'sceptic' wins).

Crucially, (d) shows us one key way in which growth could fail to be explosive, but still bring about a short TAI timeline. As such, the sceptic cannot win the debate merely by convincing us against (a) and (b); she must also convince us against (d), in favour of something like (c) or (e).

These are, of course, not exhaustive of the space of possible trajectories of capabilities progress after DRI kicks in. Moreover, the results of different trajectories of progress are also very sensitive to assumptions on the TAI threshold and the date at which DRI begins.

It seems difficult for the sceptic to deny that AI R&D automation would (at least) lead to a substantial one-time speedup in AI capabilities progress, given the potential size and speed of an automated R&D workforce.

This is especially clear under the 'automated worker' framing. According to Aschenbrenner, we should "expect 100 million automated researchers each working at 100x human speed not long after we begin to be able to automate AI research"; while Davidson estimates that "OpenAI could expand their AI researcher workforce from hundreds of experts to tens of millions."

Of course, there are serious bottlenecks to overcome in AI development, and these might become increasingly difficult to address over the next decade. But with a workforce amounting to millions more AI R&D researchers and engineers than we have today, each running faster and for longer hours than their human equivalents, it still seems likely that there would be a considerable speedup in improvements from (to borrow Aschenbrenner's phrasing) "the sheer size of the one time boost".

The upshot of this is that even without explosive growth trajectories, progress mediated by DRI *could* still be fast enough for TAI to emerge within the next ten years. Therefore, when I come to characterise a set of 'short timeline scenarios' in Chapter 3, I do not differentiate scenarios based on *whether capabilities progress is explosive or not*; instead, I differentiate them based on whether direct recursive improvement is (i) *strong enough to speed up* existing rates of capabilities progress (ii) only *strong enough to sustain* them, or (iii) *too weak to even sustain* them, in the face of bottlenecks.

Other objections

The debate of this chapter has focused on *how fast* and *how sustained* capabilities growth from recursive improvement might be. I've illustrated that there are ways for the short timeline believer to push back against the sceptic here: even taking into consideration the bottlenecks and constraints to capabilities growth highlighted by the sceptic, it appears that DRI could still, in theory, result in a trajectory of progress that is *fast enough* and *long-lasting enough* to bring about TAI within the next ten years.

However, the sceptic still has room to object on other lines. In particular: she could still question *when the recursive improvement period begins*, and how near to TAI-level capabilities we are at that time. So far in the debate, I have effectively bracketed these variables (since the arguments around them seem less interesting, and less commonly pursued in the literature), but will now consider them before ending this chapter.

When will the direct recursive improvement period begin?

The sceptic might argue that direct recursive improvement dynamics will simply not kick in any time soon – perhaps not even in the next decade – because AI R&D automation will be very difficult to achieve. In that case, DRI will not produce a short timeline. So, let's outline a few potential challenges for automating AI R&D.[95]

As elsewhere, I'm especially interested here in challenges for the development of 'automated workers', but will point out whether I think the below arguments also apply to a 'suite of services'. I'll go on to provide some brief commentary on the idea of a suite of AI services which jointly approximate the effects of automated AI R&D workers being developed in the near term. I'll note that the existence of this second plausible route to automation increases the overall likelihood that we'll see the emergence of 'AI R&D type' direct feedback loops in the near future.

Challenge (i): generality

This challenge, as it is expressed below, is clearly relevant for automated AI R&D workers. It might also apply to suites of AI R&D services.

Maybe performing AI R&D work successfully would require AI systems to already be exhibiting high levels of generality. In the sense that is most relevant here, this means not just performing well on tasks from their training distributions, but also adapting to solve new, unseen problems. After all, it does appear that success in AI R&D is, in some part, about applying previous understanding to new situations and new challenges.[96]

To advance the generality objection, the sceptic could appeal to evidence that traditional LLMs struggle to generalise (see e.g. Chollet's arguments in this direction, as outlined in Chapter 1). This might indicate that current systems are far away from the threshold for DRI.

However, arguments of this kind feel considerably weaker in the present context than they did in Chapter 1. The need for increasing the generality of AI systems was previously presented as *a barrier for developing something like AGI or HLMI* (see 'What capabilities could constitute TAI?' for definitions). We are now talking about barriers *for developing R&D AIs* – and the generality requirements on such systems seem less stringent. To the extent that these systems need to generalise successfully, this will only be for the sake of tackling unseen challenges *within* the scope of AI R&D. (By contrast, an AGI must generalise to a much wider range of unseen problems.) AI systems may very soon – if not already – exhibit sufficient generality for success in this relatively circumscribed context.

Moreover, even within the context of AI R&D, AI systems could plausibly *begin* automating significant parts of the field while exhibiting quite *low* levels of generality. In the early stages of a DRI period, these systems might achieve

[95] A line of argument I don't cover in the main text is this: even if it's easy for labs to develop systems that can automate AI R&D tasks, there will be some lag between development and deployment within those labs (due to the need for testing the systems, evaluating them against their own safety policies, and dealing with other red tape). Although this seems true, it's not clear why this lag would be significant enough to *seriously* impact the timeline to TAI. Absent the introduction of new regulatory barriers, the gap between developing and deploying automated AI R&D workers in lab settings seems likely to be to the scale of months rather than years (at least in situations where humans continue working alongside those AIs, rather than being laid off).

[96] This makes most sense in the case of automated workers. We should expect the composite system of humans-plus-AI-services in the 'suite of services' framing of DRI to be jointly be under the same generality requirements as an automated worker. And even the individual narrow components of that system *might* need to be able to generalise to new, unseen problems (provided those new problems still fall roughly in their domain of activity).

What doesn't apply to the individual components of the 'suite of services' is a requirement of *general intelligence* (in the sense of being able to perform a wide range of tasks). This is a requirement on the composite system, but not of the narrow systems comprising it. As we've discussed previously, *general intelligence* and *generality* (construed as the ability to generalise) seem to come apart.

some relatively easy wins in AI R&D by simply applying 'memorised' knowledge from vast amounts of training data related to AI R&D tasks. With such large numbers of systems applying their knowledge in the same way, and working faster and for more hours than humans, the overall effects on capabilities progress might still be significant. Of course, with each recursive improvement step, the challenges faced by R&D AIs will become harder, and there will likely be a greater need for tackling challenges which just haven't been seen before – but at the same time, the R&D AIs will be getting iteratively more capable, and likely better at generalising. Importantly: even if the need for some increased degree of generality does slow down capabilities progress at *this point*, this doesn't mean it would present a barrier for a period of DRI *starting*.

Finally, as I explained in Chapter 1: even if we are still convinced that generality is a problem here, it does seem that developers are quickly finding ways to get around the apparent generality issues facing traditional LLMs. This may be best illustrated by the breakthrough performance of OpenAI's o-series models on the ARC benchmark, which is designed to capture generality. Objections on this side have therefore lost their bite.

Challenge (ii): autonomy

This challenge is clearly relevant for automated AI R&D workers, but does not seem to apply so strongly in the case of a suite of AI R&D services.

The 'automated worker' framing of direct recursive improvement requires AIs to perform AI R&D work *largely autonomously* across a full, end-to-end workflow, without substantial reliance on humans. The sceptic might argue that we are far from developing such systems, given how reliant many frontier models are on human prompting and guidance.

However, if we consider evidence from recent AI progress, it's hard to see why this wouldn't be possible in the near term.

For example, recently, Sakana AI announced its AI Scientist system: "a fully automated pipeline for end-to-end paper generation", which they claim to be capable of independently producing machine learning research (from generating ideas, to running experiments, to generating papers, to performing peer review).[97]

Taking this at face value, it looks like AI system capabilities might at least be *nearing* a point where they could make genuine contributions to AI R&D with minimal human involvement.

This is not a *guarantee* that we'll reach this threshold in the next ten years – indeed, Sakana AI has identified many problems with the AI Scientist that will need ironing out before such models could be deployed.[98] However, it certainly seems plausible, and the sceptic has some work to do to explain why these apparent successes are not genuine evidence of being on track to do so.

Either way, it's important to note here that AI R&D automation could also be

[97] There has been some recent work to shed light on how close frontier AI systems are to automating AI R&D, conducted by METR.

[98] It's worth noting that there's considerable doubt over how big a step towards AI R&D automation Sakana AI actually represents. As Scott Alexander comments, "Is it good? Not really. Experts who read its papers say they're trivial, poorly reasoned, and occasionally make things up (the creators defend themselves by saying that "less than ten percent" of the AI's output is hallucinations). Its writing is meandering, repetitive, and often self-contradictory. [But] Like the proverbial singing dog, we're not supposed to be impressed that it's good, we're supposed to be impressed that it can do it at all."

achieved by systems with somewhat lower autonomy (say, under the 'suite of services' framing, in which humans might play a very important supervisory role). Similar direct recursive improvement dynamics might therefore be introduced in the near term even if AI systems cannot contribute to AI R&D without significant human oversight.

This motivates a closer look at the near-term plausibility of developing a 'suite of services' whose effects on AI R&D might approximate those of an automated worker.

Automated workers vs suite of services

A suite of AI R&D services might be somewhat easier to develop than automated AI R&D workers.

Given how I originally told the 'suite of services' story (in the section titled 'Direct recursive improvement'), the value of the individual AI systems here primarily lie in their capacity to be instructed by human workers to produce certain outputs, which humans would subsequently utilise. Because of this, each component system of a suite of services probably would not need to exhibit as high a degree of autonomy as an automated worker would.

(If this is correct, it might mean that the feedback loops introduced by a suite of services are weaker than those that would be introduced by automated workers, since humans are apparently acting as more of a bottleneck to AI progress in the 'suite of services' story. However, since my framing of 'automated workers' also permits some degree of human involvement in AI R&D – albeit of a less critical kind – this difference is not likely to have *dramatic* effects on the comparative speeds of capabilities improvements in both cases.)

Secondly, although each component of a suite of services might need to possess some degree of *generality* (in the sense of being able to tackle unseen problems), they would not individually need to perform well on as wide a range of problems in AI R&D as automated workers would. Put differently: each component system here will be far less *generally intelligent* than an automated AI R&D worker.

The key requirement on each of the component systems in a suite of services is simply being highly skilled on some narrow range of AI R&D tasks – and this is something we're already seeing good examples of. For example, AI is already being used for chip design, data generation, writing code, and building applications.

The remaining challenge for developing a suite that would approximate the effects of an automated AI R&D worker is building a *sufficiently comprehensive set* of individual highly-skilled systems. Jointly, these systems need to automate all or most aspects of AI R&D, and we're certainly not there *yet*. It might also turn out that certain tasks are much harder to automate than others, and that we've already plucked most of the low-hanging fruit here (for example, maybe tasks like writing code and generating data are just *far* easier

to automate than, say, generating R&D ideas which are genuinely novel).

I won't go into this any further here. For now, I'll just note that developing narrow systems which skilfully perform the most difficult AI R&D tasks would still probably be easier than developing an automated worker (which would individually need to perform all of the most difficult AI R&D tasks, *and more*).

This is not to say that a comprehensive suite of AI R&D services is likely to be developed before an automated AI R&D worker – this might largely come down to where the most R&D investment and effort goes, and it does seem that major labs are at least *trying* to build more broadly-skilled agents. Instead, the point I am making is this: the existence of this second (and possibly less demanding) strategy for introducing 'AI R&D type' direct feedback loops *strengthens the case for DRI emerging in the near future*. If for whatever reason, automated workers cannot be developed in the next few years, developing a suite of AI R&D services might still be a possibility – and might have quite similar effects.

As a final, related point here: suppose there is some subset of tasks in AI R&D which are unusually difficult to automate, and this means that in the near term, we are neither able to build a *truly end-to-end* automated AI R&D worker nor a *truly comprehensive* suite of AI R&D services. In these cases, we could still see partial automation of the field – and this might still be enough to set off recursive dynamics which speed up AI capabilities progress to an extent. I explore this idea briefly below.

The dynamics of AI R&D automation

In Situational Awareness, Aschenbrenner argues that the work involved in AI R&D is not that complicated to replicate with AI systems. Indeed, if we take trendlines of AI capabilities growth seriously, AIs might be capable of doing most of the relevant tasks within the next few years:

> "the job of an AI researcher is fairly straightforward, in the grand scheme of things: read ML literature and come up with new questions or ideas, implement experiments to test those ideas, interpret the results, and repeat. This all seems squarely in the domain where simple extrapolations of current AI capabilities could easily take us to or beyond the levels of the best humans by the end of 2027"

Aschenbrenner does, however, admit that the automation of AI R&D tasks might exhibit long tails, since "the last 10% of the job of an AI researcher might be particularly hard to automate". Still, even partial automation of the field – whether it's just the last 10% of tasks that must be supplemented by humans, or far more – would still kick off some process of direct recursive improvement. In this instance, we might just expect *weaker feedback* from each improvement step to the next (and, as a result, slower corresponding growth in capabilities) than in a 100% (or near-100%) automation scenario, where human performance is not really acting as a serious bottleneck. In light of the reflections under 'Who wins the tug of war?', it seems that this is still compatible with a short timeline.

In fact: it seems that the most plausible story for DRI is one that begins with a period of partial automation of AI R&D *which gradually transitions into full automation*, rather than a sudden jump from 0% to 100% automation. In the early stages of partial automation, the 'AI R&D type' direct feedback loops in play are relatively weak in effect or restricted in scope – but with each improvement to the capabilities of R&D AIs, more of the field can successfully be automated, and these feedback loops thus get stronger and stronger.

This idea has been built into the most prominent arguments for DRI, such as the characterisations from Drexler and Tom Davidson. This results in a more gradual curve of AI capabilities progress as the field becomes increasingly automated, rather than a sudden spike in capabilities at the moment some high percentage of automation is reached, but needn't significantly damage the case for short timelines, as the variety of trajectories shown in Figure 2.9 should help to make clear.

How far away is TAI when direct recursive improvement begins?

Perhaps the threshold for TAI is just too far away for us to reach within the next ten years, even with fast direct recursive improvement dynamics in play.

Here, the sceptic could appeal to some of the arguments made in the previous chapter on compute scaling: for example, that the compute requirements for TAI are extremely high, or that the current paradigm is fundamentally unable to scale to TAI.

However, these arguments are somewhat harder to make in this context, given the apparent bottleneck-breaking capacity of direct recursive improvement dynamics. Put differently: making a convincing argument of this kind is harder to do here, because with the introduction of direct feedback loops in the believer's story, drivers of capabilities progress have the potential to be much stronger than seems feasible in compute scaling stories *without* DRI.

For example, if the (effective) compute threshold for TAI is very high, we can now argue that having a huge automated AI R&D workforce driving improvements in hardware performance and algorithmic efficiency will enable us to meet it quickly. Likewise, if the current limitations of traditional LLMs are an obstacle for reaching TAI-level capabilities, we can now argue that a huge automated workforce would rapidly find solutions to overcome these limitations or develop a new paradigm.

What this means, effectively, is that counterarguments of this kind would need to be made *in combination with* some of the other counterarguments of this chapter (i.e. that DRI-mediated capabilities growth could not be extremely fast, or could not be sustained for very long, or will not start any time soon) if it is to cast very serious doubt on recursive improvement as a pathway to short timelines.

Chapter 3: Short TAI timeline scenarios

The previous chapters of this report laid out and examined the debates over two possible mechanisms of fast AI capabilities progress – compute scaling and recursive improvement – and their potential to produce TAI within the next ten years. Through this, several compelling arguments for short TAI timelines have already emerged, and have been shown to stand up reasonably well against some of the key sceptical arguments.

I now go on to synthesise the core argumentative threads of these chapters to generate a set of scenarios which exhibit short TAI timelines. Each of these scenarios is underpinned by different assumptions about capabilities progress – and each seems, in light of previous reflections, to represent a plausible future for the development of AI. Through mapping out this space of scenarios, a more robust case for believing in short TAI timelines is brought into focus.

This approach is inspired by Convergence Analysis' research agenda for *AI Clarity*, which emphasises scenario planning as a tool for exploring and addressing AI risk under uncertainty. My specific choice of methodology here, which distils key ideas from earlier arguments under a set of scenarios, has two purposes: (a) to clarify how different assumptions about the world could support a short TAI timeline, and (b), to begin building a more concrete picture of what the next ten years of AI development might actually look like in a short timeline world.

Chapter roadmap

First, I characterise five distinct scenarios in which AI capabilities progress is primarily driven by some combination of compute scaling and/or recursive improvement. These five scenarios differ on the values they assign to a series of parameters, as represented in Figure 3.1 and Table 3.2.

I then consider whether any important pathways of progress which are plausible and compatible with short TAI timelines have not been adequately represented within that list. Here, it is noted (with François Chollet as a leading example) that one can construct stories of TAI arriving within the next decade which neither rely heavily on the continued success of compute scaling with the current paradigm, nor on the emergence of direct recursive improvement dynamics. These stories typically involve some change in approach to AI R&D which, once adopted, brings us very close to TAI. Two additional scenarios are outlined on this basis.

In all, the seven resulting scenarios are:

(1) **'Straight Path'.** Compute scaling just works.

(2) **'Rising Tide'.** IRI breaks bottlenecks.

(3) **'New Spark'.** Moderate DRI sustains progress.

(4) **'New Engine'.** Strong DRI accelerates progress.

(5) **'Dual Engine'.** Joint compute scaling + DRI accelerates progress.

(6) **'LLM Hybrid'.** Hybrid AI systems are the trick for achieving TAI.

(7) **'Intelligent Network'.** Networks of AI systems are the trick for achieving TAI.

From the reflections of this chapter emerges a strengthened case for short TAI timelines: it seems that such timelines are compatible with, and robust to, a variety of different assumptions about the world.

I end this chapter with some reflections on the strategic relevance of *which scenario we are in*, given a short TAI timeline.

Five scenarios based on compute scaling/ recursive improvement

From the arguments of previous chapters, five plausible scenarios with short TAI timelines can now be constructed. I generate these via a decision process which is represented by the scenario tree below (Figure 3.1).

The scenario tree

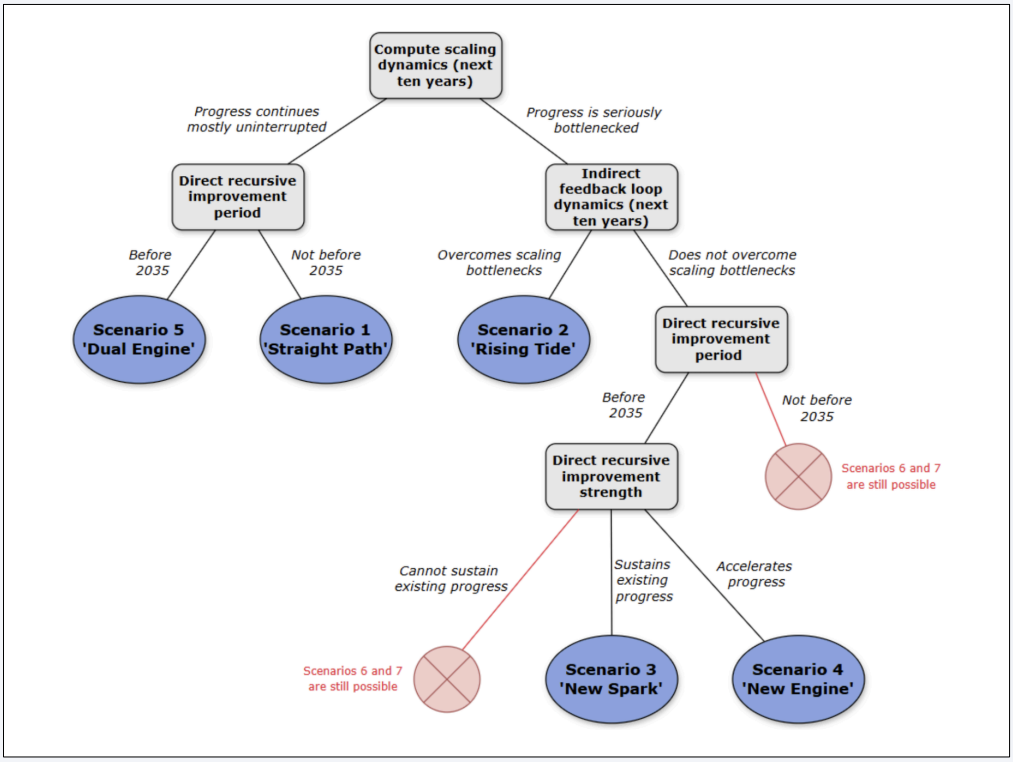


Figure 3.1: The scenario tree. A full-sized version of this figure can also be found in Appendix A.

The rest of this section will be focused on outlining the methodology for this decision process in more detail, zooming into the parameters of progress represented by each of the nodes of the above tree, and describing the five resulting scenarios.

I begin (in 'What matters most?') by roughly introducing the four dimensions on which the chosen compute scaling/recursive improvement scenarios vary. This will be followed by an explanation of the methodology for generating the scenarios, descriptions of each of the five resulting scenarios, and a summary table of the assumptions behind each scenario. Additional methodological details can be found in Appendix B.

What matters most? A rough pass introduction to the scenario methodology

The timeline to TAI depends upon an extremely complex web of variables; this much is evident from the previous chapters of this report. However, in order to construct a suitably focused set of scenarios, I want to hone in on the most crucial building blocks of the arguments we have seen so far for short timelines.

Simplifying accordingly, I think what *really matters* to the debate – what the arguments for short TAI timelines hinge on – is as follows:

(i) **Whether compute scaling continues** to happen and yield strong results in capabilities improvements over the next ten years, or instead, this route of progress **gets seriously bottlenecked** by something (be it data, paradigmatic limitations, power requirements, the supply chain, or anything else).

(ii) **How strong** the effect of **indirect feedback loops** will be on AI capabilities progress. These feedback loops are already in effect, but it's uncertain how powerful they will be e.g. against bottlenecks to compute scaling.

(iii) Whether **direct recursive improvement will begin** in the next ten years. Although there are some signs which indicate that current systems are *moving towards* the capabilities necessary for automating significant parts of AI R&D, it's not clear whether they will actually reach this point within the next ten years.

(iv) If direct recursive improvement does begin, **how strong** the effects of this will be on capabilities progress. For example, it's uncertain how the trajectory of direct recursive improvement will be affected by bottlenecks to progress, or how useful these recursive dynamics will be in overcoming such bottlenecks.[99]

Scenario generation methodology

To generate short timeline scenarios which differ on important assumptions, I seek to capture a range of distinct positions on the set of questions above.

[99] Note that the 'strength' of DRI might be best understood as a combination of variables (1) and (2) from the previous chapter. That is:
(1) How fast the capabilities improvements from DRI are
(2) How long this period of capabilities growth is sustained for
These two variables were difficult to separate out in the debate of Chapter 2, and have accordingly been combined in this chapter.

First, I reframe each of questions (i)-(iv) as a 'key parameter', representing a critical dimension along which stories of the next decade of AI development might vary. For each parameter, I define parameter values which capture distinct states that dimension of future progress could be in.

These parameters and their corresponding values form the basis of the scenario tree shown earlier (Figure 3.1). The tree is structured to guide us through a systematic process of making assumptions on the points (i)-(iv).

The answer to an additional question, "*how far away is TAI?*", is then implicitly taken to be "*as close as it needs to be for the assumptions made on (i)-(iv) to result in a short timeline*" in all cases where this further assumption seems plausible. (Only two pathways are excluded on the basis of implausibility here.)

As a result, the end state of each pathway through the tree (with just two exceptions) is a distinct short TAI timeline scenario that is both **plausible** and **compatible with the assumptions** made along the way.

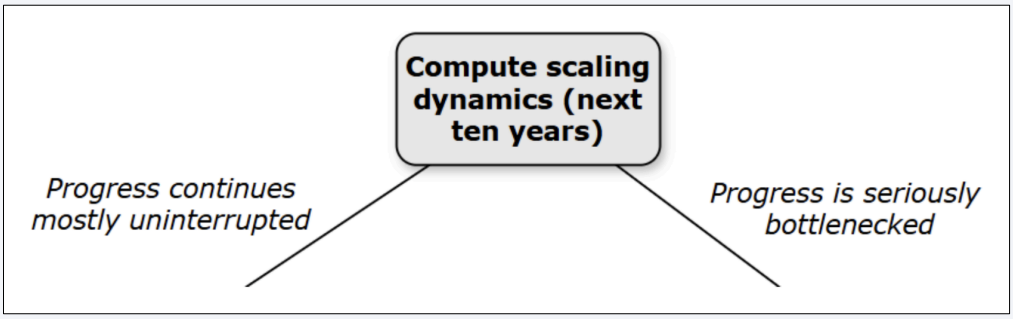
The key parameters, and the values they can take, are outlined below. Screenshots of the relevant nodes of the scenario tree are also included alongside these descriptions.

Key parameters and parameter values

Question	Key parameter	Parameter values
(i)	Compute scaling dynamics (next ten years)	Progress continues mostly uninterrupted / Progress is seriously bottlenecked
(ii)	Indirect feedback loop dynamics (next ten years)	Overcomes scaling bottlenecks / Does not overcome scaling bottlenecks
(iii)	Direct recursive improvement threshold	Before 2035 / Not before 2035
(iv)	Direct recursive improvement strength	Cannot sustain existing progress / Sustains existing progress / Accelerates progress

Table 3.1: Table summarising key parameters and parameter values.

(i) Compute scaling dynamics (next ten years)
Parameter values: *Progress continues mostly uninterrupted / Progress is seriously bottlenecked*.



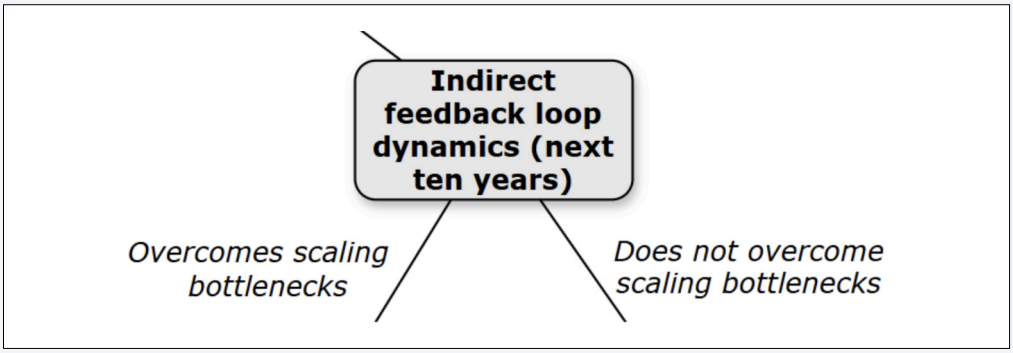
One important determinant of the world we are in is the fate of the current trajectory of capabilities progress, driven by compute scaling. Over the next

ten years, compute scaling could either *continue* to drive fast improvements in AI capabilities, more or less as before, or else get *seriously bottlenecked* on something.[100] By the latter option, I mean that one or both of the following outcomes is realised:

- Compute is overtaken by some other factor of AI development (such as data quality/quantity, or some specific limitation of the current paradigm) as the main bottleneck for capabilities progress, such that the gains from compute scaling plateau; or
- AI labs are unable to keep accessing or training systems on enough compute, such that current trends in compute growth break down.

I do not further differentiate these two outcomes under the 'Progress is seriously bottlenecked' parameter value, since both cases are likely to have a (roughly) similar impact on the plausibility of short TAI timelines.

(ii) Indirect feedback loop dynamics (next ten years)
Parameter values: *Overcomes scaling bottlenecks / Does not overcome scaling bottlenecks*.



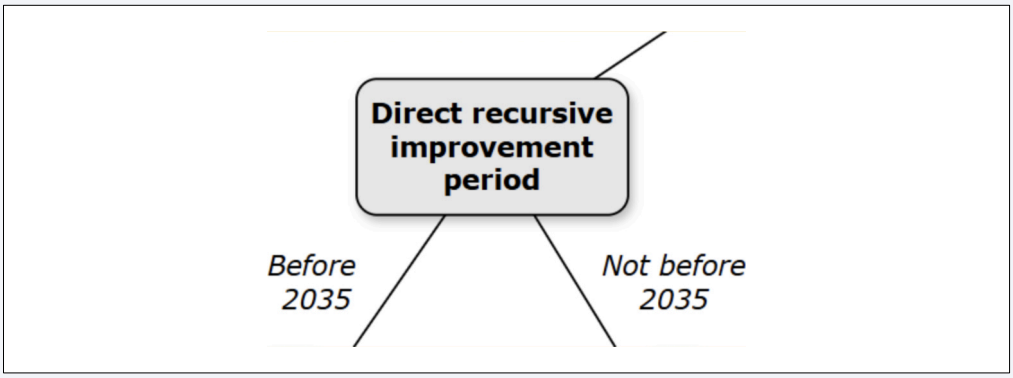
If the compute scaling pathway does get 'seriously bottlenecked' over the next ten years, it is relevant to ask whether 'indirect' feedback loops (whether economic, scientific, or political) will eventually gain enough traction to *overcome* those bottlenecks to compute scaling.

If scaling bottlenecks are overcome in this way within the next ten years, the previous trajectory of capabilities progress would thus resume, and the plausibility of a short timeline would be increased; if this does not happen, then something else will likely be necessary to take us to TAI by 2035.

[100] I acknowledge that this distinction might be better understood as a whole spectrum of outcomes, corresponding to varying degrees of slowdown to the current trajectory of compute scaling. However, for simplicity in differentiating a short list of scenarios, I treat it as a binary threshold. I make similar simplifications when selecting values for the other key parameters.

(iii) Direct recursive improvement period

Parameter values: *Before 2035 / Not before 2035*.

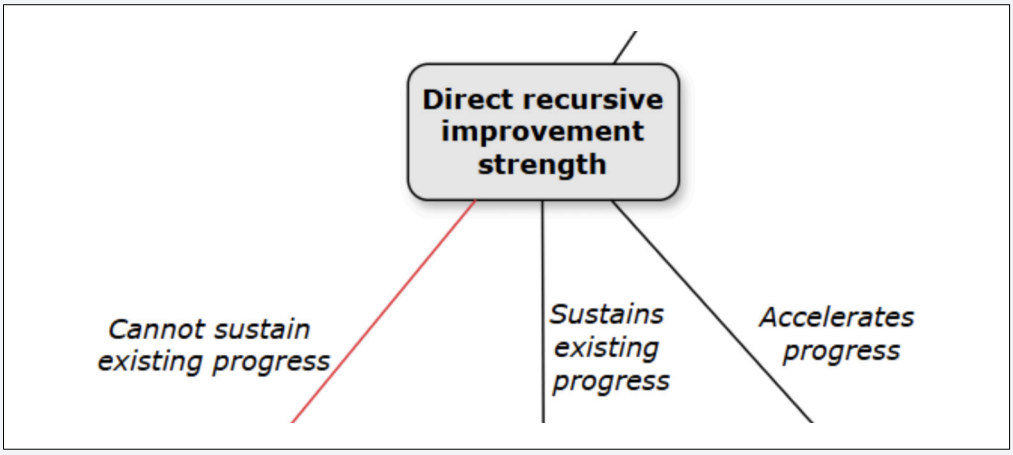


Unlike indirect feedback loops, which are already supporting AI progress to some degree, the emergence of direct feedback loops would mark a meaningful *new* development in the field that could define the next era of capabilities progress.

Therefore, regardless of whether compute scaling gains are continuing uninterrupted or encountering significant bottlenecks, we should ask whether a period of direct recursive improvement will kick in *within the next ten years*[101] (either as a reinforcement or replacement to continued gains via compute scaling). This might not happen – a period of direct recursive improvement could begin *after 2035, or never*. But if it does happen, the trajectory of AI progress could be transformed, perhaps boosting the chances of a short TAI timeline.[102]

(iv) Direct recursive improvement strength

Parameter values: *Cannot sustain existing progress / Sustains existing progress / Accelerates existing progress*.



In Chapter 2, I illustrated how the emergence of direct recursive improvement dynamics in AI development could result in a variety of different trajectories of capabilities progress. The trajectory that is realised depends on the strength of the positive feedback loops in comparison to the constraints at each improvement 'step', the increasing effects of certain bottlenecks to recursive

[101] As I explained in Chapter 2, it's not very likely that there will be a sudden shift from 0% AI R&D automation to 100% AI R&D automation. When I say that *a period of DRI begins by 2035*, I do not necessarily mean that 100% automation of AI R&D has happened by this point; a lower percentage of automation could have roughly similar effects on the overall outcomes for AI progress.

Here, I do not define the *minimum* level of automation which I would take to satisfy the claim that DRI has begun; I simply treat this as some unspecified percentage. Of course, if I were building a probabilistic model based on this scenario tree, the value of this percentage would be a key input. But in the present context, it only affects the likelihood of different parameter values at and downstream of this node in the tree (i.e. the likelihood of DRI beginning in the next ten years, and the probabilities of different strengths of DRI). Since I'm not dealing with probabilities in this report, I am content to leave this unspecified.

Moreover, and like with (i), the value of this parameter might actually be better understood as a spectrum rather than a binary distinction. An alternative parameter to consider might therefore be '% of AI R&D automation by 2035'.

[102] Whether it does or does not provide this 'boost' in likelihood depends on the value of parameter (iv).

improvement, and the rate of diminishing returns.

There are also synergies here with the dynamics of compute scaling leading up to, and during, the recursive improvement period. In a world where compute scaling has come up against significant bottlenecks and the existing trajectory of AI capabilities progress is at risk of breaking down / has already broken down, the effects of a period of direct recursive improvement could be any of the following:

- DRI may be too weak to overcome these bottlenecks and therefore *unable to sustain*/restore previous rates of progress;
- DRI may be strong enough to overcome these bottlenecks, and thereby *able to sustain* or restore previous rates of progress, but not accelerate them; or
- DRI may be more than strong enough to overcome these bottlenecks, and thereby *able to accelerate* progress.
    - An 'acceleration' of capabilities progress could look like a one-time speedup or a period of sustained acceleration (and the latter option could itself correspond to a variety of different growth modes). However, I do not differentiate 'acceleration' any further here, since, as noted in 'How do you win the tug of war?', these options could have similar consequences for the plausibility of a short timeline.

Notes on scope and structure of the scenario generation

It is possible to prompt a different set of value assignments on *all four* of the above parameters along each pathway through the tree, but this would result in 24 (=2*2*2*3) distinct pathways, each capturing a slightly different future. This level of granularity is not necessary for our purposes.

Firstly, not every 'future' here is even *coherent*: for example, it does not make sense to ask what the strength of DRI is if a period of DRI has not actually begun in this ten year time frame.

Secondly, not every future is *meaningful* in the context of this report. For example, suppose that compute scaling is seriously bottlenecked over the next ten years, but indirect feedback loops gain sufficient traction to overcome those bottlenecks, restoring the previous rates of progress under compute scaling. In this case, it doesn't add much value to then consider whether a period of DRI *also* begins before 2035; the two possible outcomes of this additional question would very closely resemble the two corresponding outcomes of the pathway on which compute scaling had continued uninterrupted in the first place. (That is: there's no need to further differentiate Scenario 2 based on whether DRI begins before 2035; those outcomes are already basically captured by Scenario 1 and Scenario 5).

To avoid unnecessary complexity, I limit the selection of parameter 'prompts' in the tree (which are represented as nodes) to those which either yield a *coherent* and *meaningfully new* scenario, or otherwise result in a 'dead end'. As

seen in Figure 3.1, this results in a refined set of five short TAI timeline scenarios and two dead ends.

These methodological details on the scope and structure of the tree are expanded upon in Appendix B.

Scenario descriptions

This decision process, as represented by the scenario tree in Figure 3.1, generates five short TAI timeline scenarios which have been named as follows:

- **Scenario 1:** 'Straight Path'
- **Scenario 2:** Rising Tide'
- **Scenario 3:** 'New Spark'
- **Scenario 4:** 'New Engine'
- **Scenario 5** 'Dual Engine'

These five scenarios are each described briefly below. Their assumptions on each guiding question of the scenario tree are also summarised in Table 3.2.

SCENARIO 1

'Straight Path'. *Compute scaling just works.*

Compute scaling with the current paradigm continues to yield results and does not become *seriously* bottlenecked in the next ten years.[103] There are problems to solve along the way (e.g. on the side of data or algorithms), but there are quick fixes available (e.g. synthetic data generation[104] works well, and unhobbling leads to easy improvements in LLM generality). Direct recursive improvement does not kick in at any point, but doesn't need to; compute scaling is enough to produce TAI by 2035.

SCENARIO 2

'Rising Tide'. *IRI breaks bottlenecks.*

Compute scaling gets seriously bottlenecked on something in the next ten years (e.g. at some point, developers just can't afford enough compute to continue scaling systems up). However, indirect feedback loops in the background gain traction over the next ten years. (For example, AI systems attract some capital which can be reinvested into procuring more compute, the scaled-up AI systems perform better and attract even more capital, and so on.) This helps to lift capabilities progress out of a plateau. Direct recursive improvement *could* also kick in at some point, but doesn't need to; compute scaling plus indirect recursive improvement is enough to produce TAI by 2035.

SCENARIO 3

'New Spark'. *Moderate DRI sustains progress.*

Compute scaling gets seriously bottlenecked on something in the next ten years. Indirect feedback loops do not gain sufficient traction to lift capabilities

[103] Or, if it does get seriously bottlenecked, another form of compute scaling (e.g. with run-time compute rather than training compute) works just fine. I don't mention this option explicitly in my scenarios, but take it to basically be a variant of what I call 'compute scaling' here. Of course, it only applies in cases where the bottleneck to compute scaling is not a *lack of physical compute*.

[104] Recall from Chapter 2 that I do not consider synthetic data generation alone as sufficient for underpinning what I call a period of 'direct recursive improvement'. I do, however, accept that AIs which generate data could bring about a much more restricted (and therefore weaker) form of the same dynamic.

progress out of this plateau. However, a period of direct recursive improvement soon kicks in. It's strong enough to sustain current rates of capabilities progress. Systems are near enough to TAI-level capabilities at the time that direct recursive improvement kicks in for TAI to be produced by 2035.

SCENARIO 4

'New Engine'. *Strong DRI accelerates progress.*

Compute scaling gets seriously bottlenecked on something in the next ten years. Indirect feedback loops do not gain sufficient traction to lift capabilities progress out of this plateau. However, a period of direct recursive improvement soon kicks in. It's strong enough to accelerate capabilities progress. (For example, there could be a one-time step change in the rate of capabilities progress, or a sustained period of continuous acceleration.) Even if systems are far away from TAI-level capabilities at the time that direct recursive improvement kicks in, this doesn't matter; direct recursive improvement leads to such fast (and/or prolonged) capabilities progress that TAI is still produced by 2035.

SCENARIO 5

'Dual Engine'. *Joint compute scaling + DRI accelerates progress.*

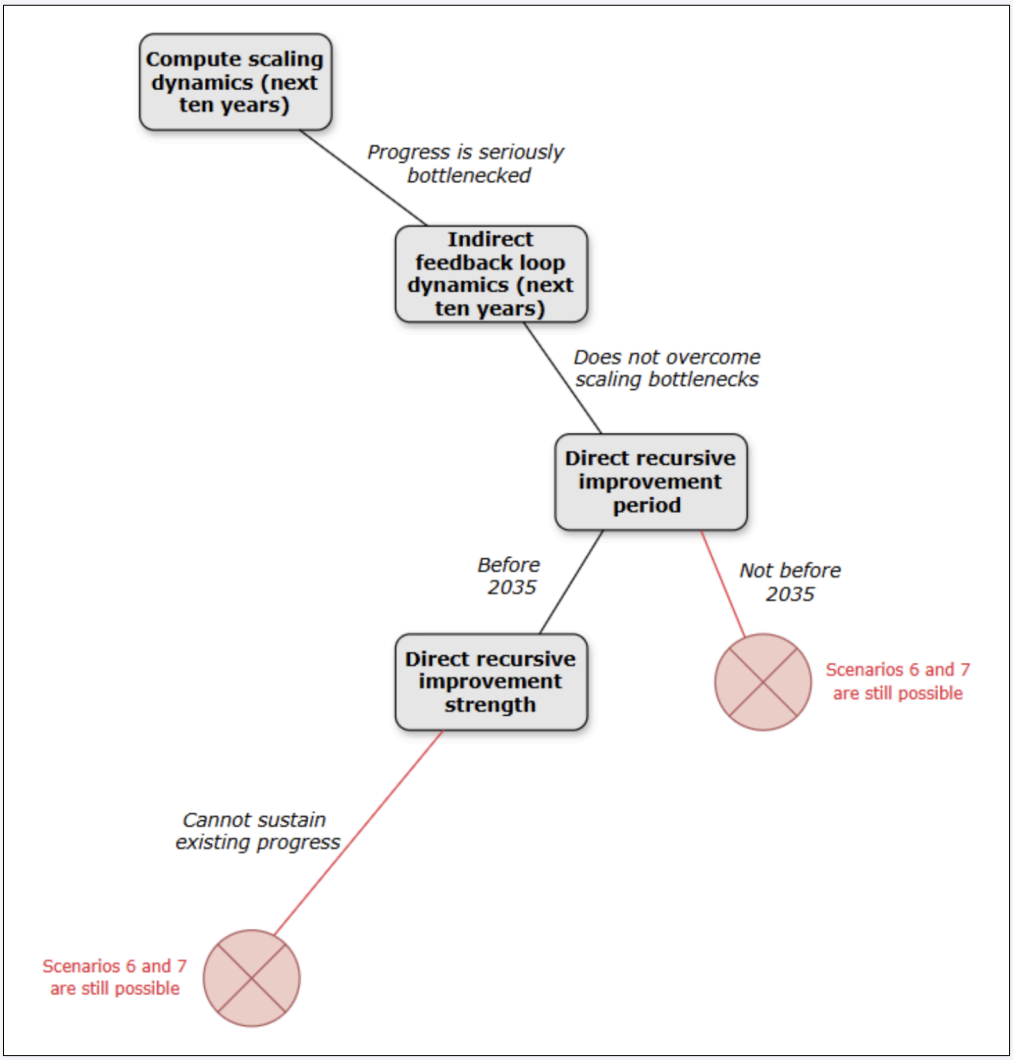
As in Scenario 1, compute scaling with the current paradigm continues to yield results and does not become seriously bottlenecked on anything in the next ten years. There are problems to solve along the way, but there are quick fixes available. Direct recursive improvement also kicks in within the next ten years. Even if systems are far away from TAI-level capabilities at the time that direct recursive improvement kicks in, this doesn't matter; direct recursive improvement plus continued compute scaling leads to such fast (and/or prolonged) capabilities progress that TAI is still produced by 2035.

Table of scenario parameter values

Parameter values for each short timeline scenario				
Scenario	*Compute scaling dynamics (next ten years)*	*Indirect feedback loop dynamics (next ten years)*	*Direct recursive improvement period*	*Direct recursive improvement strength*
1	Continues uninterrupted	N/A	Not before 2035	N/A
2	Seriously bottlenecked	Overcomes scaling bottlenecks	N/A	N/A
3	Seriously bottlenecked	Does not overcome scaling bottlenecks	Before 2035	Sustains progress
4	Seriously bottlenecked	Does not overcome scaling bottlenecks	Before 2035	Accelerates progress
5	Continues uninterrupted	N/A	Before 2035	N/A

Table 3.2: Parameter values for each of the first five short timeline scenarios.

A note about the 'dead ends' of the scenario generation process. There are two pathways through the scenario tree (Figure 3.1) in which both compute scaling and recursive improvement are, in some important sense, blocked. These two pathways are isolated in the screenshot below.



Given the combinations of assumptions made on these two decision pathways, there is nothing to go off in either case which indicates that AI capabilities will improve very much over the next decade.[105] But there are stories we could tell, invoking dynamics *besides* compute scaling and recursive improvement, that could explain how we could still get TAI by 2035 in these cases. Some of the available options here are discussed in the next section ('Have we missed anything important?') and captured under two additional scenarios.

Have we missed anything important?

Having outlined this set of short timeline scenarios, it's worth asking: what does this list leave out?

The five scenarios above capture several plausible pathways to TAI by 2035, but they are not exhaustive of the whole landscape of possibilities. Rather, they illustrate how various combinations of assumptions about compute scaling and recursive improvement – assumptions that are hotly debated in the literature – can be consistent with a short TAI timeline.

However, there are also conceivable stories in which TAI arrives by 2035 that do not significantly rely on assumptions about future compute scaling or the emergence of recursive improvement dynamics. These alternatives demonstrate that even if one is sceptical of the arguments presented in the previous chapters, a belief in short TAI timelines can still be justified.

In this section, I highlight a few of these alternative pathways of progress, and capture them under two additional short timeline scenarios. By broadening the scope of possibilities in this way, the case for short TAI timelines becomes more robust (as will be explained in 'The case for short TAI timelines is strengthened').

I begin by noting François Chollet's perspective as a specific example of an alternative view: it does not clearly mesh with any of the five scenarios described above, but still takes short TAI timelines as a serious possibility.

François Chollet's alternative view

Chollet, the sceptic?

So far in this report, Chollet has come across as a sceptic. Indeed, many of the sceptical arguments that I have detailed in previous chapters, against the plausibility of short TAI timelines via compute scaling or via recursive improvement, have been drawn from his work.

Chollet's actual perspective on timelines is worth clarifying here. At the very least, it seems clear that he has historically been sceptical of the two following, specific claims:

(1) The current paradigm (of 'traditional LLMs'[106]) can be scaled with increased compute to TAI by 2035.

(2) A period of direct recursive improvement can take us to TAI by 2035.[107]

[105] And it wouldn't be in good faith to construct a short timeline scenario here by just assuming that we're *already* brushing up against TAI-level capabilities, such that almost *no* progress is required to get us there; this assumption feels quite implausible.

[106] Recall from Chapter 1 that I use this term roughly to refer to models like GPT-3 and -4, in direct contrast to (for example) OpenAI's o3 model. The outputs of 'traditional LLMs' are largely determined by an underlying transformer-based neural network trained on next-token prediction. The o3 model appears to diverge from this basic structure, as it (probably) features *something* built on top of an underlying neural network that makes *substantive contributions* to the model's outputs via programme search. Accordingly, Chollet seems more hopeful about capabilities progress in the direction of o3, as will become clear later on in this section.

[107] Chollet's scepticism might actually be restricted to a more narrow claim than (2): he might believe that DRI based on *traditional LLMs* cannot result in TAI by 2035, but DRI involving another type of system could. (This is not absolutely clear from his writing on the subject.) This is a minor point which has little bearing on this section, or on the validity of the final scenario selection.

As a result, he would likely be sceptical of all five of the scenarios presented earlier in this chapter. These scenarios all heavily rely on some combination of compute scaling and/or recursive improvement. In addition, the pathways which involve significant contributions from compute scaling are best interpreted with reference to the current paradigm of traditional LLMs; indeed, my original framing of 'the compute-centric scaling hypothesis' made specific reference to the scaling up of transformer-based neural network architectures with deep learning.

Surprisingly, however, it turns out that Chollet is not really *a sceptic of short TAI timelines in general*. In September 2024, he tweeted that he believes AGI "is in fact likely in the next 10-15 years". He noted that his conception of AGI is not strictly in keeping with some other conceptions that have been in play in the debate (that is to say, it's "not an artificial human mind"). However, he still seems to be referring to something that is plausibly transformative.

His view therefore points to the existence of some *other* viable pathway for developing TAI within the next ten years. What is it?

Chollet, the believer?

While Chollet has doubts that the current deep learning paradigm of traditional LLMs can become capable enough to constitute AGI (even with significant scaling or recursive improvement), he has higher hopes for a hybrid paradigm.

Specifically, he has argued[108] that a system combining deep learning with discrete programme search (DPS) could achieve considerably better performance on a wide range of tasks than traditional LLMs do. The major limitations of deep learning (e.g., as outlined in Chapter 1, limitations on the generality of traditional LLMs) could be supplemented by DPS, since one has strengths where the other has weaknesses. In fact, Chollet effectively views these two methods as a means of replicating different important modes of thinking: deep learning is a good fit for 'system 1' thinking, while DPS is useful for 'system 2' thinking (under Kahneman's definitions).

Adopting a hybrid approach which effectively combines these two modes of thinking in a single AI system might bring AI capabilities far closer to something we could reasonably call TAI. Moreover, implementing this approach might be achievable within the next few years.

In fact, the very recent development of OpenAI's o3 model (which, Chollet suggests, "represents a form of *deep learning-guided program search*") appears to *already* be a step in this direction[109] – and a promising one, with the model having achieved groundbreaking performance on the ARC benchmark which many traditional LLMs have struggled to contend with.

There are lots of unknowns here. At the time of writing, the role played by programme search in o3 has not actually been confirmed by OpenAI. And if Chollet's suggestions about the architecture of o3 *are* accurate, it's still not clear whether this means that OpenAI developers are now shifting their efforts towards something like a hybrid paradigm, nor is it clear whether further work

[108] See e.g. timestamp 0:49:35 in his interview with Dwarkesh Patel.

[109] It's not *exactly* what Chollet was originally talking about (though it certainly seems similar). Chollet himself comments that: "There are however two significant differences between what's happening here [with o3] and what I meant when I previously described "deep learning-guided program search" as the best path to get to AGI. Crucially, the programs generated by o3 are *natural language instructions* (to be "executed" by a LLM) rather than *executable symbolic programs*. This means two things. First, that they cannot make contact with reality via execution and direct evaluation on the task – instead, they must be evaluated for fitness via another model, and the evaluation, lacking such grounding, might go wrong when operating out of distribution. Second, the system cannot autonomously acquire the ability to generate and evaluate these programs (the way a system like AlphaZero can learn to play a board game on its own.) Instead, it is reliant on expert-labeled, human-generated CoT data."

in this direction would even continue to yield impressive results. However, the existence and recent successes of o3 at least indicate that Chollet's vision of improved capabilities via some kind of hybrid paradigm might not just be far-future speculation.

Other Chollet-like views

Chollet's story is not focused on some mechanism for AI capabilities progress which continually drives improvements over the next ten years. Instead, it falls under a category of stories in which, in the near future, there is some *change in approach to AI development* – some new trick or clever idea – which, once adopted, brings us much closer to TAI.

These stories don't *preclude* the possibility of continued compute scaling or a period of DRI contributing to capabilities improvements over the next decade, but they do point to some new approach to AI development as the *primary explanation* for the arrival of TAI. Since the scenario tree from Figure 3.1 is silent on whether any major labs adopt a substantial (and successful) change of approach within the next ten years, these stories are not explicitly represented under the five scenarios I characterised previously.

Possible 'tricks' for AI development

Hybrid LLM paradigms. In Chollet's story, the 'trick' is to develop a hybrid AI system combining LLMs with DPS. One could also tell a variety of similar stories which involve equipping LLMs with different forms of symbolic reasoning or new learning methods. These stories can be seen as characterising different forms of 'hybrid LLM paradigms'.

Other algorithmic innovations? More broadly, there are many stories we could tell where the 'trick' that gets us to TAI is some algorithmic innovation (say, a new way to implement recurrence or improve chains of thought). This wouldn't necessarily have to involve a hybrid system; instead, we might just be looking at a more efficient (but still 'traditional') LLM. However, if the resulting improvement in AI capabilities is largely due to this innovation suddenly unlocking much higher levels of *effective compute*, I view the story to be more or less a variant of Scenario 1 'Straight Path' – it's basically another way in which the current paradigm could scale to TAI, with compute.

A meaningfully distinct scenario might be found here if the algorithmic innovation in question represents something that could reasonably be called a *paradigm shift*. While it's unclear exactly what this would have to look like, I use the term 'paradigm shift' to loosely refer to any fundamental departure from the current dominant framework: transformer-based neural networks trained using next-token prediction.[110] A shift away from this setup would, in my view, result in systems that cannot be best described merely as more efficient or advanced versions of LLMs.

I believe that the kind of paradigm shift which could most *plausibly* be achieved within the next ten years would reap some of the existing benefits of

[110] Under this rough definition, and assuming that Chollet is correct about o3's architecture, o3 appears to be *at least* a step in the direction of a paradigm shift – if not already a new paradigm. Although it is *guided* by a transformer-based neural network trained using next-token prediction (specifically, by a GPT), it seems that there are *substantive contributions* to its outputs from add-ons to this base architecture, in a sense that seemingly sets it apart from 'traditional LLMs'. I discussed this in more detail in Chapter 1.

transformer-based neural networks/deep learning/next-token prediction, enhanced by the introduction of other complementary architectures or techniques; this is why I focus here on 'hybrid LLM paradigms' in particular.

Networks. Another 'trick' for developing TAI within the next decade involves the composition of multiple distinct AI systems in the form of a network. (This could, for example, look similar to Drexler's Comprehensive AI Services model; though note that Drexler himself expects direct recursive improvement to drive the development of advanced AI systems, and his personal view might therefore be better understood as a variant of Scenario 3, 4, or 5.)

Although major developers like OpenAI have recently been focused on creating systems that are increasingly general, it's plausible that TAI need not actually take the form of a single, unified, general-purpose system. Instead, it could emerge from the combined capabilities of several specialised systems working in concert, each highly skilled in its own domain, and together comprising a highly general system.[111] This would potentially sidestep some of the sceptical challenges levelled in the previous chapters of this report concerning the generality of traditional LLMs, which afflict stories of both compute scaling and recursive improvement.[112]

Since we already have AI systems that are performing significantly above human-level in their narrow domains, such as AlphaFold, it seems plausible that a network of this kind could be composed within the next decade.

This trick might be viewed as another form of hybrid architecture, since it's likely that a network of this kind would compose LLMs with other systems (especially if developed within the next ten years). In this case, instead of a single hybrid system, we would have a *hybrid network*. However, the real 'trick' in this story is specifically the introduction of the network; this is the crucial change in approach that takes us from a disparate collection of very narrow systems towards something genuinely transformative.

Two new scenarios

Based on the above, I propose extending our list of short timeline scenarios with the two following additions:

- **Scenario 6: 'LLM Hybrid'.** A hybrid architecture is developed which combines LLMs with a form of symbolic reasoning or new learning methods. This displays much higher levels of generality than the current paradigm. Relatively minor or fast improvements to this hybrid paradigm are sufficient to achieve a form of TAI by 2035.
- **Scenario 7: 'Intelligent Network'.** Before 2035, many systems, each with narrow capabilities, are composed together in a network (e.g. in the style of Drexler's Comprehensive AI Services). The combination of these systems' individual capabilities constitutes a genuinely transformative composite system.

Unlike the first five scenarios in this chapter, which represent discrete

[111] The main distinction between this scenario and the 'suite of AI R&D services' I described in Chapter 2 is that the latter is scoped to the field of AI R&D, while a network of the kind I am presently describing would cover a very broad range of economically relevant tasks (in order to add up to something like AGI/human-level machine intelligence).

[112] Here and in reference to Chollet's story, I have highlighted generality/general intelligence as important features of TAI. However, possessing *very high levels of generality* or being *fully general* may be too high a bar here; AI systems (or networks of AI systems) could be genuinely transformative while falling short of these conditions. Moreover, there are other features of some relevance to the question of whether a system could transform society, such as its flexibility or adaptiveness. Ultimately, achieving something like general intelligence is seen by many in this debate as a sufficient condition for TAI, and is often a focus of the discourse around timelines. Since an analysis of all different possible features or forms of TAI is out of scope, I stick to the example of generality / general intelligence here.

pathways through a scenario tree based on differing assumptions, these two scenarios do overlap (since a network of intelligences could include, or instantiate, a kind of hybrid system).

Relationship to the previous scenario framework

These two scenarios do not fit neatly into the scenario tree given earlier in this chapter. This is because, at the level of detail currently presented, they are basically silent on the dynamics of compute scaling and recursive improvement over the next ten years. As a result, they are essentially compatible with *any* set of assumptions we might make on the parameters which underpin that framework (unless these parameters are construed as *strictly* referring to progress on systems within the current paradigm).

Notably, this means that Scenarios 6 and 7 provide viable pathways to short TAI timelines even in the 'dead end' cases of the scenario tree (highlighted in red in Figure 3.1), where compute scaling is seriously bottlenecked in the next ten years, indirect recursive improvement isn't strong enough to overcome these bottlenecks, and direct recursive improvement is either weak or absent.

Implications

The case for short TAI timelines is strengthened

The existence of a plurality of routes through which TAI could feasibly be achieved by 2035 is noteworthy, and should strengthen our overall degree of belief in short timelines.

Through this chapter, it is evident that short TAI timelines are compatible with a variety of different assumptions about the future. In particular:

- Scenarios 1-5 are based on meaningfully different assumptions about scaling and recursive improvement dynamics over the next decade. Taken together, this set of scenarios illustrates how if one route of capabilities progress is slow or begins to plateau, other mechanisms could soon kick in through which TAI might still quickly be achieved.
- Scenarios 6-7 characterise pathways to short TAI timelines which do not significantly rely on either of these mechanisms. So, even if both compute scaling and recursive improvement are unsuccessful, or are less powerful drivers of progress over the next ten years than expected, there are other directions through which TAI might still be achieved by 2035.

Of course, we certainly can't make *any* combination of assumptions we like and still end up with a short TAI timeline. For example, depending on the parameter values selected at each node of the scenario tree, one has to make a compatible assumption on *how far away TAI is* in order to arrive at one of the first five short timeline scenarios.

It's also worth highlighting that Scenarios 6 and 7 are somewhat speculative. In both cases, I haven't examined the arguments for and against their plausibility,

drawn out the assumptions that they rely on, or even referenced sources which do this in any detail. While o3 *may* provide some supporting evidence for the ideas behind Scenario 6, not enough is known about OpenAI's latest releases to draw out any solid conclusions about the likelihood of this scenario here. I've included Scenarios 6 and 7 in this chapter mainly on the basis of their prima facie appeal, with the intention of illustrating the breadth of options that could be available to the believer in short TAI timelines.

I therefore don't put too much stock in any *one* particular scenario being realised (and hold some scenarios in a more cautious regard than others). But I do view this diverse set of scenarios, taken together, as having some evidentiary weight: *to deny short TAI timelines, you have to say no to a lot of seemingly plausible assumptions about the next ten years of AI development*.

Which scenario?

Although I don't put too much stock in any one scenario, it does *matter*, from a strategic perspective, which short timeline scenario we are in. Amongst other things, this influences the likely values of the following parameters.

- The type / strength of TAI systems to expect in 2035.
- The likely trajectory of capabilities progress up to 2035.
- The warning signs to expect (if any) on the path to TAI.

The relationships between these parameters and the scenarios described in this chapter are not rigid; each scenario leaves room for the details to be fleshed out in a variety of ways. However, in each scenario, certain combinations of parameter values will be more *likely* than others. In turn, the relative likelihoods of different risks, as well as the appropriate methods of governance, vary across them.

Future work could be done to unpick these relationships and thereby improve our understanding of the risks of short TAI timelines. For now, I offer only some **very speculative thoughts** about the parameter values that seem likely in each of the scenarios discussed in this chapter.

If progress is driven solely through compute scaling, with no recursive improvement (Scenario 1) we might imagine that:

- *Type of TAI*: The TAI systems that arrive by 2035 are effectively scaled up versions of current neural network-based systems.
- *Trajectory*: There is largely predictable progress leading up to the arrival of TAI, closely corresponding to increases in (effective) compute. The absence of major bottlenecks results in a fairly smooth, uninterrupted trajectory.
- *Warning signs*: We see near-TAI systems before we see TAI. A series of warning signs also comes from leading labs acquiring vast amounts of compute and running increasingly large training runs.

If recursive improvement, or joint compute scaling + recursive improvement, accelerates capabilities progress (Scenarios 4 and 5) we *might* imagine that:

- *Type of TAI*: The TAI systems which have emerged by 2035 are superintelligent, due to accelerated progress over the next decade. They look very different to current systems, having benefited from the innovations of a greatly expanded AI R&D field.
- *Trajectory*: There is a step change or period of accelerating capabilities growth once the relevant recursive improvement dynamics kick in.
- *Warning signs*: A new era of AI capabilities progress is heralded by the arrival of AI systems which can automate significant parts of AI R&D. There is not much time to act on this warning sign, as TAI follows shortly thereafter.

By contrast, if recursive improvement just helps to sustain current trends of capabilities progress (Scenarios 2 and 3) we *might* imagine that:

- *Type of TAI*: The TAI systems which have emerged by 2035 are below-superintelligent, since current progress rates have only been sustained over the next decade. Architecturally, they look similar to current systems.
- *Trajectory*: Due to scaling bottlenecks, there is a temporary slowdown of progress. This is followed by a restoration of previous rates of progress once the relative recursive improvement dynamics pick up.
- *Warning signs*:
  - In Scenario 2, we observe an uptick of 'indirect' feedback loops (e.g. through race dynamics or increased investment) before we see TAI.
  - In Scenario 3, we see the arrival of AI systems which can automate significant parts of AI R&D before we see TAI.
  - In both cases, there is time to act on this warning sign; we see near-TAI systems before we see TAI.

In alternative scenarios (Scenarios 6 and 7), we *might* imagine that:

- *Type of TAI*: The TAI systems that emerge by 2035 do not look like scaled up traditional LLMs. Instead, they look like hybrid LLM systems or a distributed network of narrow intelligences.
- *Trajectory*: There is a discontinuous jump in capabilities or step change in the rate of capabilities growth once an effective new 'trick' to AI development is deployed.
- *Warning signs*: A new era of AI capabilities progress is heralded by the adoption of some 'trick'. There is not much time to act on this warning sign, as TAI follows shortly thereafter.

Further exploration of the short TAI timeline scenarios presented in this report – *what they might look like*, and *what we should do about them* – is highlighted in the Conclusion as a potential priority for future research.

Taking stock

In characterising the seven scenarios of this chapter, I have illustrated how a variety of different assumptions about the world could plausibly support a short TAI timeline, reinforcing the case for believing in them. This exercise has also helped us to begin building a more concrete picture of what the next ten years of AI development might actually look like in a short timeline world (though I acknowledge the need for further exploration here, given the variability of strategic implications across these worlds).

The insights from this scenario analysis, taken in combination with the arguments of previous chapters, equip us to better engage with the broader debate over short timelines and the complex body of evidence that underpins it – a task I will now take up in the Conclusion.

Conclusion

In this report, I have explored routes through which AI capabilities progress could be fast enough to reach transformative levels within the next ten years. Across Chapter 1 and Chapter 2, I conducted a detailed examination of two key mechanisms for AI capabilities progress, evaluating a variety of arguments for and against their yielding a short TAI timeline. In Chapter 3, I combined core threads of these arguments in different ways to produce a list of seven distinct scenarios with short TAI timelines.

What has emerged from this is a complex picture of driving factors and constraining factors for AI capabilities progress (and an equally complex push-and-pull between a 'believer' and a 'sceptic') – but it's one in which, ultimately, short TAI timelines end up seeming very plausible. In particular, as I'll argue below:

- The existence of a plurality of pathways to short TAI timelines, each based on meaningfully different assumptions, increases the likelihood of a short TAI timeline being realised; and
- The body of evidence in support of short TAI timelines is rapidly growing as the AI field progresses.

After covering these points, I will summarise the difficulties faced by the sceptic of short TAI timelines in her debate against the believer, highlight key uncertainties that remain, and point to possible directions for future work.

There are a plurality of pathways to short TAI timelines

There are several different routes through which TAI could conceivably be achieved by 2035, each based on meaningfully different assumptions about the world and the future.

Through the debate between the believer and the sceptic which occurs across Chapter 1 and Chapter 2 of this report, it becomes clear that a belief in short timelines is compatible with a variety of different background assumptions (for example, about scaling, the current paradigm, and the strength of different drivers and restraints of AI capabilities progress). In fact, as I argued in 'Who wins the tug of war?', we can even believe in short TAI timelines while agreeing with the essence of many of the sceptic's points, and incorporating some effects of her posited restraints of AI capabilities progress into our models.

The diversity of possible pathways to short TAI timelines is reflected (to some extent) within the set of scenarios outlined in Chapter 3. One of these short timeline scenarios is a straightforward story of compute scaling; others point

to ways that a combination of compute scaling and recursive improvement can enable continued/accelerated capabilities progress; and others don't even really rely on either of these mechanisms in any significant way.

These scenarios, taken together, provide a strong body of evidence for the plausibility of TAI arriving by 2035. They illustrate ways in which, if capabilities progress on one pathway is slow or begins to plateau, other mechanisms could soon kick in through which TAI might still quickly be achieved.

It should be noted that this pathway-shifting would not just happen coincidentally; rather, the AI labs that are seeking to develop TAI will actively try to move in a new direction (say, towards a different scenario on the list from Chapter 3) if their existing one isn't yielding the results, or the pace, they are hoping for. For example, they could choose to deploy AI systems to automate parts of their R&D efforts, kickstarting a type of direct recursive improvement. They could also ramp up the effect of economic feedback loops by pushing hard for greater investment/higher profits, and reinvesting a larger proportion of this into capabilities R&D.

In summary, the plurality of pathways for producing TAI by 2035 means that:

(i) There are a variety of combinations of assumptions under which we can reasonably believe in short timelines to TAI; and

(ii) The failure or implausibility of any *one* pathway to a short TAI timeline is not strong evidence that a short TAI timeline will not be achieved.

Of course, there are also many pathways through which short TAI timelines would *not* be realised. This doesn't have much bearing on my above argument; I'm not claiming that we have more reason to believe in short TAI timelines than the alternatives, but simply that short TAI timelines are *plausible*.

Under the next heading, I'll argue a little more directly that there does, in fact, appear to be a growing imbalance between the evidence for short TAI timelines and the evidence against it, in favour of the former. I don't take this argument as far as concluding that short TAI timelines are more likely than not, for reasons I elucidate later on under 'Areas of uncertainty'.

There is a growing body of evidence pointing towards short TAI timelines

When I set out to write this report in May 2024, the body of evidence I was contending with felt very mixed. In fact, as I was reading, my own opinions were continuously shifting back and forth along a spectrum of scepticism and belief, and were overall only slightly skewed in the direction of shorter timelines. If you'd asked me then what the conclusions of this report might be, I would probably have supplied more middle-of-the-road answers.

Over the course of writing, however, I witnessed a series of developments in

the AI field and the AI governance community, the majority of which seemed to be reinforcing the likelihood of short TAI timelines. In the last few months, this included (but was not limited to):

- **OpenAI's report on its o1 model**, which indicates the potential for making significant AI capabilities improvements via *another* form of compute scaling (namely, scaling run-time compute) where overhangs from training can potentially be exploited. In addition, o1's improved performance on the ARC benchmark in comparison to GPT-4 illustrates the promise of chain-of-thought reasoning as a means of improving LLM generality (which helps to address a major point of contention from the sceptic of compute scaling). – *September 2024*.
- **The release of OpenAI's o3 model**, which reinforces the theories about run-time compute and chain-of-thought reasoning based on o1. Not much else is known about o3 – but speculatively, this model might also represent a step away from 'traditional LLMs', towards something like a hybrid paradigm. If sceptics are correct to claim that something like a paradigm shift will be necessary for developing TAI, o3 *might* be seen as evidence that such a shift is actually achievable within the next decade. Relatedly, the successes of o3 (e.g. on the ARC benchmark) have provided some evidentiary support for the plausibility of Scenario 6. – *December 2024*
- **Epoch's *Can AI Scaling Continue Through 2030?* report**, which predicts that "by 2030 it will be very likely possible to train models that exceed GPT-4 in scale to the same degree that GPT-4 exceeds GPT-2 in scale", since proposed 'bottlenecks' to compute scaling are not likely to have a serious effect over the next five years. – *August 2024*.
- **Sakana's announcement of its AI Scientist model**, "a fully automated pipeline for end-to-end paper generation". If Sakana's claims are taken at face value[113], this may be seen as evidence that the field is nearing a point where AI systems could start making genuine contributions to AI R&D, and a period of direct recursive improvement could thus begin. – *August 2024*.
- Microsoft CEO **Satya Nadella** effectively claiming that a phase of recursive improvement has already begun ("we are using AI to build AI tools to build better AI"). – *October 2024*.
- The publication of prominent, evidence-backed research falling firmly on the side of short TAI timelines, such as **Aschenbrenner's *Situational Awareness***. – *June 2024*.
- **François Chollet**, one of the most prominent 'sceptical' voices on AI capabilities progress of recent years, clarifying that he does in fact believe TAI is "likely in the next 10-15 years" (i.e., it appears that he takes short timelines to be a serious possibility). Many of the arguments against compute scaling and recursive improvement that I have presented in this report, and which are popular in the literature, have drawn from Chollet's writings. It seems that some of the most compelling sceptical arguments

[113] Recall from footnote 98 that there is substantial debate over whether the results of this model point to a major development in the field, or have in fact been seriously overblown.

in this debate don't actually commit us to very strong (or very general) scepticism of short timelines after all.[114] – *September 2024*.

- Another apparent 'sceptic', **Yann LeCun**, revealing that he expects human-level AI within the next ten years (even though he thinks current systems are below cat-level intelligence). – *October 2024*.[115]

Recent evidence against short timelines? For a fair comparison, we must now ask: which *recent* developments in the field point in the opposite direction? In general, it's hard to answer this question, because it's not always clear what a development that supports longer timelines would be, or how we would spot it. After all, AI systems *continuing to be unable to do some task* isn't really a 'development' in the way that successes in capabilities improvements are, and isn't likely to be reported as energetically. Moreover, opinion pieces that hark the imminent arrival of ultra-powerful, society-transforming AI systems may be more likely to garner attention than those which simply say "well, it looks like we still have some way to go".

One recent development which *has* been interpreted as pointing towards shorter timelines came out after this report was substantially written. Specifically: in November 2024, news broke that OpenAI's unreleased GPT-5 model represents a smaller leap forward in AI capabilities than its predecessors did. Given the timing of this news, the implications of this for the timelines debate have only been lightly discussed in the present report (see, for example, 'What evidence is there for the scaling hypothesis?' in Chapter 1). The issue is worth briefly revisiting here, and engaging more closely with in future work.

Some commentators have speculated that the news about GPT-5 indicates that gains from scaling training compute are plateauing, and have taken this to herald a "big AI slowdown". At the current time of writing (December 2024), there's simply not enough information to determine whether what we're seeing is actually a breakdown of scaling laws, or just a temporary blip in GPT progress. But if scaling laws are breaking down, the believer can still make a case for short timelines:

- As I argued in Chapter 1, recent evidence from OpenAI's o-series models illustrates the potential for gains through scaling run-time compute, even if the gains from scaling training compute do diminish.
- As I argued in the previous section, the plurality of short TAI timeline scenarios outlined in Chapter 3 illustrates that even if all forms of compute scaling break down in the near term, there are other routes of progress through which TAI could still be achieved by 2035.

Still, it would be misleading to suggest that *all* recent developments have pointed towards shorter timelines. There have been, and there might continue to be, developments in the field that the sceptic can point to in defense of her claims.

And in fact, there's more to say in the sceptic's favour here. For example, despite some major AI developments in recent months, the following things do

[114] As explained in Chapter 3, Chollet is *genuinely* sceptical of the specific claim that "compute scaling will take the current paradigm to TAI by 2035", as well as the specific claim that "recursive improvement will produce TAI by 2035". However, he's not a sceptic of short timelines *in general*, as his tweet (linked in the main text) illustrates.

In this sense, Chollet's position exemplifies the *plurality of pathways to short timelines* I discussed earlier. Even if you're a firm sceptic of achieving short timelines via one or both of the key mechanisms outlined in this report, you need not be a firm sceptic of short TAI timelines being achieved *at all*.

[115] Several additional developments of note happened in late January 2025, in the period of a few weeks between this report being finalised and its publication. For example:

- Executive orders on AI and power announced by the Trump administration, alongside a new $500 billion project with OpenAI, (all of which I have noted in footnotes in Chapter 1) indicate a new push to rapidly develop large-scale, advanced AI systems, and a commitment to removing any barriers to doing so. This means that some of the main conceivable bottlenecks to scaling up frontier models (on the side of investment, infrastructure, and so on) may soon be overcome or dissolved. – *20th-21st January 2025*
- The arrival of open source Chinese AI model DeepSeek, which is comparable in performance to its US competitors, might also contribute to faster capabilities progress – whether this is through other labs accessing its

source code, or as a result of race dynamics between the US and China. However, I haven't engaged with this news item in any detail. – *Most discourse on this has been in late January 2025*

- Dario Amodei (Anthropic) seemingly called for developing AI systems which can recursively improve, in response to news about DeepSeek. – *29th January 2025*

[116] A key example of this is how (as I noted earlier in the main text) the believer can respond to claims that gains from scaling training compute are diminishing by pointing to the recent news about OpenAI's o-series models, which suggests that significant capabilities improvements could instead be made by scaling run-time compute.

remain true:

- There are still many economically relevant tasks that current AI systems cannot do; it seems that some areas of capabilities progress are still hampered, to an extent, by the limitations of current systems or external constraints on their development. (See, e.g. the results of a recent study of LLM capabilities by Apple engineers.)
- (Perhaps relatedly) There are still many prominent thinkers who are *unambiguously sceptical* of short TAI timelines, such as Robin Hanson.

Despite this, I still think that recent evidence is imbalanced in favour of short timelines. Evidence from the last few months of AI progress hasn't been completely one-sided. However, it is true that the list of apparent successes in the field has been growing remarkably rapidly, and is becoming difficult to ignore.

Importantly, as this list of successes grows, it provides more and more routes through which the believer can respond to any points which seem to be in favour of the sceptic. As Aschenbrenner notes in *Situational Awareness*, claims made by sceptics (specifically, those claims of the form "LLMs will not be able to do X within the next Y years") keep getting falsified with each new success story that comes out. And we have seen that, for every challenge that the sceptic has argued that LLMs currently face, or will potentially face – whether this is diminishing returns from compute scaling, running out of data, being unable to generalise to unseen problems, or the need to master certain tools – there are some recent developments in the field that the believer can point to which suggest this challenge could soon be met.[116] The equivalent doesn't *quite* seem to hold true for the sceptic who is trying to maintain her position.

On balance, I do think there is a growing asymmetry of evidence here: over time, the debate appears to be shifting overall towards short timelines as a likely outcome of capabilities R&D efforts.

The prospect of short timelines cannot be ignored

How should one react to this debate?

My arguments so far suggest that there is substantial evidentiary weight behind the believer's arguments, and that it's becoming increasingly difficult for the sceptic to maintain her stance against short TAI timelines. To do so, she must not only counter the multitude of possible pathways to TAI by 2035, but also respond to this rapidly growing body of evidence which appears (in large part) to be in conflict with her claims.

This doesn't mean that it is unreasonable to doubt short TAI timelines, or that the sceptic has failed to identify any serious barriers for AI developers to overcome on the path to TAI. On the contrary, I think that most of the challenges for capabilities progress highlighted in this report are very real, and

will have a tangible impact on the trajectory of progress over the coming years.

What this *does* mean is that the prospect of TAI arriving within the next ten years cannot be dismissed. Firstly, *none of the sceptic's arguments seem decisive* against short TAI timelines; the believer generally has room to maintain that the mechanisms for capabilities progress she has appealed to will be strong enough to mitigate the effects of the identified restraints, or to overcome bottlenecks. Secondly, none of the challenges highlighted in this report are obviously insurmountable in the next ten years *with sufficient effort and investment* over this period. And, importantly: AI labs are *trying* to produce TAI, and they want to get there quickly. They might just succeed.

Given the above, and given the implications for AI governance of facing short timelines, decision-makers should treat these scenarios as genuine possibilities, and prepare accordingly.

What now? Key uncertainties and future research directions

Areas of uncertainty

I believe this report has provided good evidence to support the claim that *short TAI timelines are plausible*. However, I am reluctant to take this conclusion any further (for example, to make the stronger conclusion that *short TAI timelines are more likely than not*). This is because of several areas of remaining uncertainty.

The first and most important of these is methodological, and ultimately a consequence of the limited scope of this report: I have not conducted the same level of analysis of the prospect of *long TAI timelines* (i.e., timelines greater than ten years) as I have for that of short timelines. My examination of the debate was fairly balanced on both sides for the first two chapters of this report – but in Chapter 3, I significantly furthered the case for short timelines, explicating the range of different assumptions which could plausibly be compatible with a short timeline, and capturing these under a robust set of scenarios. By contrast, I did not construct any equivalent set of 'long TAI timeline scenarios', or work through the range of assumptions which could be compatible with *them*.

Given this, I am reluctant to go as far as weighing the overall likelihood of short timelines against the likelihood of long timelines; doing so would require making a direct comparison that I do not believe this report has licensed me to make.

Other key uncertainties at this stage directly concern the content of this report's arguments. Below, I highlight a few uncertainties which seem to be of particular importance at the time of writing (December 2024):

Could the compute scaling era be over, and what does this mean for future

AI progress? Are we really seeing a plateau in gains from scaling training compute, already? If so, attention might soon shift towards scaling run-time compute (in combination with other strategies). How effective will this be? Can the existing pace of capabilities improvements be sustained in this way? And can we effectively 'stack' gains from increased training compute *alongside* those from increased run-time compute?

Will there be a 'gap' between the compute scaling and direct recursive improvement eras? In Scenarios 3-5 on my list, I (implicitly) assume that there won't be any gap between the compute scaling and direct recursive improvement eras; that is, before we essentially run out of juice from scaling, AI capabilities will already have reached a point at which AI systems can start making significant contributions to the field, and DRI can begin. But this is not a given. To determine how plausible these scenarios really are, we need to understand:

- (As above) Is the compute scaling era already coming to an end? If so, capabilities progress might slow down substantially before a DRI threshold has even been reached, making Scenarios 3-5 seem unlikely.
- What are the most likely bottlenecks at the point at which compute scaling breaks down, and what *specific capabilities* would an automated AI R&D workforce actually need in order to overcome them?
- What are the dynamics of automating AI R&D likely to look like? For example: how gradual is automation? What is easy to automate, and what is hard? Does the distribution of *ease of automation* exhibit 'long tails'?

How should we understand the results from OpenAI's o1 and o3 models? As I have speculated in multiple places in this report, the apparent successes of these new releases might have significant bearing on the likely pathway and trajectory of future AI development. However, there's currently little information publicly available about these models, and they also haven't been around for long. As a result, I haven't been able to fully work through the consequences of these developments on the overall timelines debate.

A note on some areas for expansion

In Chapter 3, I outlined a set of scenarios in which, by 2035, TAI has emerged. Despite uncertainties, I maintain that it is *plausible* that we are in one of these scenarios.

However, as I highlighted in 'Which scenario?', it's not just the timeline to TAI that is of strategic importance for the AI safety and governance community: it's also important to consider the pathway of progress we might end up taking to get there. Features of this pathway influence the type of transformative system that arrives first, the trajectory of capabilities growth leading up to its arrival, how far we will have surpassed TAI by (if at all) in 2035, how quickly we might subsequently progress to even more powerful forms of AI, and the warning signs we can expect to have (if any) along the way.

As such, it is not just useful to understand *whether* we might be in a short TAI timeline scenario; we also need to consider *which one(s)* we are most likely to be in, and what this means from the perspective of safety and governance. This motivates future work on better understanding the short TAI timeline scenarios presented in this report – their respective likelihoods, ways to determine which scenario we are in, what the risks look like in each case, and how we should approach them. My speculations at the end of Chapter 3 about the state of the world in each scenario can be seen as an initial, but *very tentative*, step in this direction.

A 'bounty list' for future research

A broad list of potential future research directions is provided below. Some of these research questions aim to address the uncertainties described above, and some seek to expand my analysis on the specific scenarios presented in this report.

I believe all of these topics are worthy of further exploration. The Convergence Analysis team may pursue some selection of these in future work. However, the team cannot cover all of this ground, and would therefore strongly welcome other researchers to use any of the below as prompts for building on the work done in this report.

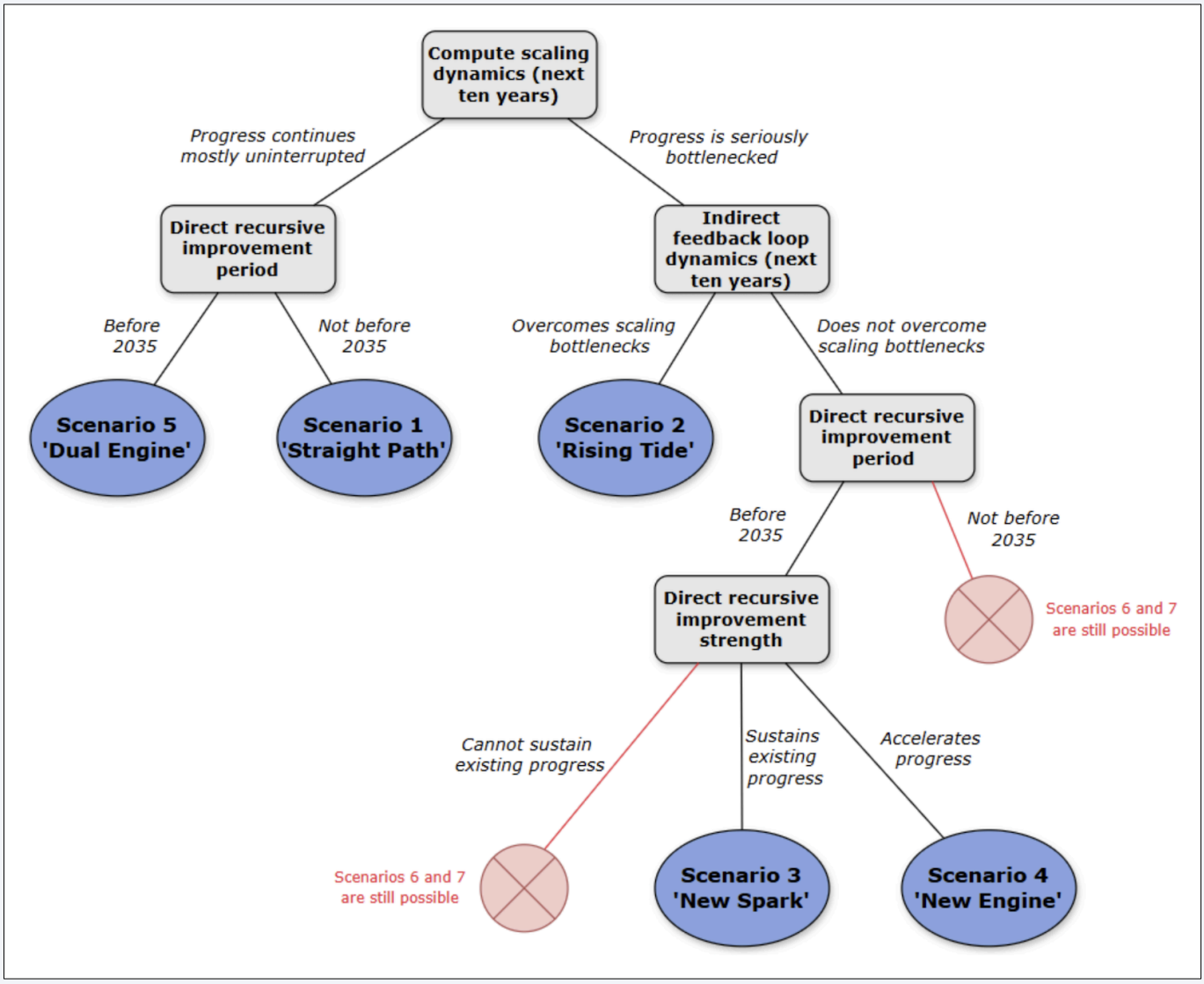
POTENTIAL RESEARCH DIRECTIONS

- Further exploration of the short timeline scenarios presented in this report:
  - Which are most likely?
  - Will there be warning signs for being in such a scenario? If so, what will they look like?
  - What does the world look like in 2035?
  - What would the strategic implications of being in such a scenario be?
  - What's missing from this list of scenarios?
- What does the development landscape look like in short timelines?
  - Which players could develop the first TAI?
  - How competitive is the landscape?
  - How great a lead does the first TAI developer have on others?
  - Soft nationalisation?
- 'Patterns' of AI progress:
  - What are the different possible trajectories for AI capabilities

progress?
    - What would cause them?
    - What seems likely?
    - What are their strategic implications?
- What are the different ‘worldviews’ behind different AI timeline predictions?
- Limitations of current/traditional LLMs and their implications for AI timelines:
    - Generality
    - Adaptive learning
- Are performance gains from compute scaling plateauing? What new evidence would suggest that scaling laws are indeed breaking down, and how strongly would this affect the case for short timelines?
- How should we understand o1 and o3 in this picture: how do they work, what can we learn from them, and what are the likely next steps from here?
- Understanding ‘general intelligence’ and the capacity for intelligent systems to break bottlenecks
- Automation of AI R&D
    - What capabilities do AIs need for automating AI R&D? What challenges will they need to overcome to do this?
    - What will the dynamics of automation look like? Which tasks will be automated first? How long from partial automation to full automation? What are the impacts on the rate of R&D returns?
- What would an intelligence explosion look like, and what would its strategic implications be?
- What will diffusion look like once transformative AI is developed? Could the timeline to transformation be much longer than the timeline to transformative AI?

If you have any opinions on Convergence Analysis' research priorities or would like to collaborate, please get in touch with the team at research@convergenceanalysis.org. You can also contact the author of this report, Zershaaneh Qureshi, via email or on LinkedIn.

Appendix A: Scenario tree (full page version)



Appendix B: Further methodological details

Recall that the scenario tree of Figure 3.1 (repeated in Appendix A) was generated on the basis of a set of parameters and possible parameter values. A summary of these parameters and parameter values is repeated here below, for reference:

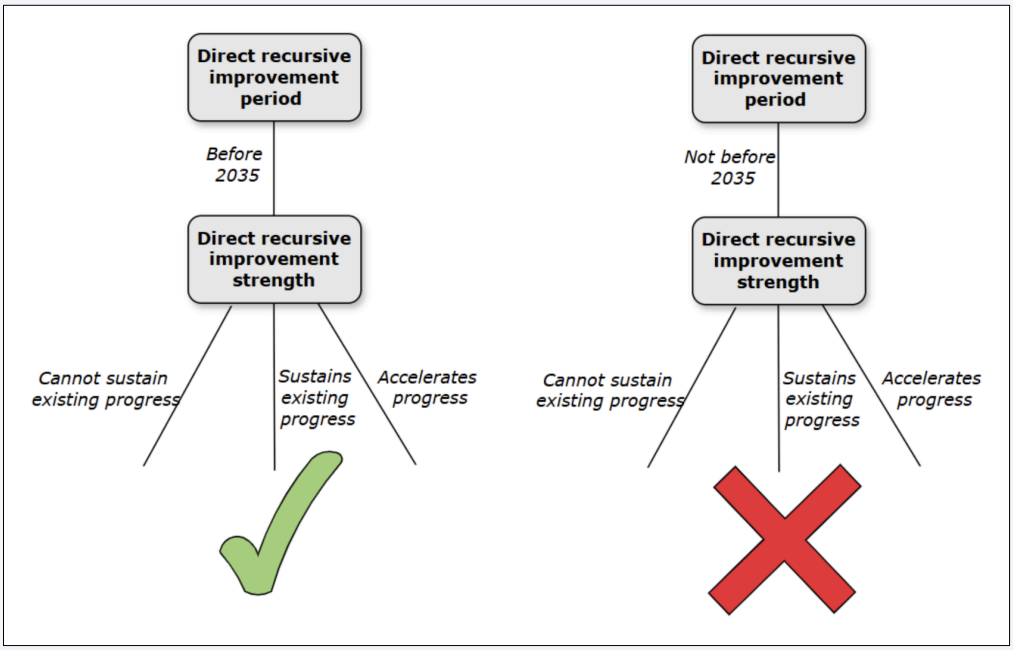
Question	Key parameter	Parameter values
(i)	Compute scaling dynamics (next ten years)	Progress continues mostly uninterrupted / Progress is seriously bottlenecked
(ii)	Indirect feedback loop dynamics (next ten years)	Overcomes scaling bottlenecks / Does not overcome scaling bottlenecks
(iii)	Direct recursive improvement threshold	Before 2035 / Not before 2035
(iv)	Direct recursive improvement strength	Cannot sustain existing progress / Sustains existing progress / Accelerates progress

Notes on the scope and structure of the scenario generation process

It is possible to prompt a different set of value assignments on all four of these parameters along each pathway through the tree, but this would result in 24 (=2*2*2*3) distinct pathways, each capturing a slightly different future. This level of granularity is not necessary for our purposes.

Firstly, not every 'future' here is even coherent: for example, it does not make sense to ask what the strength of DRI is if a period of DRI has not actually begun in this ten year time frame.

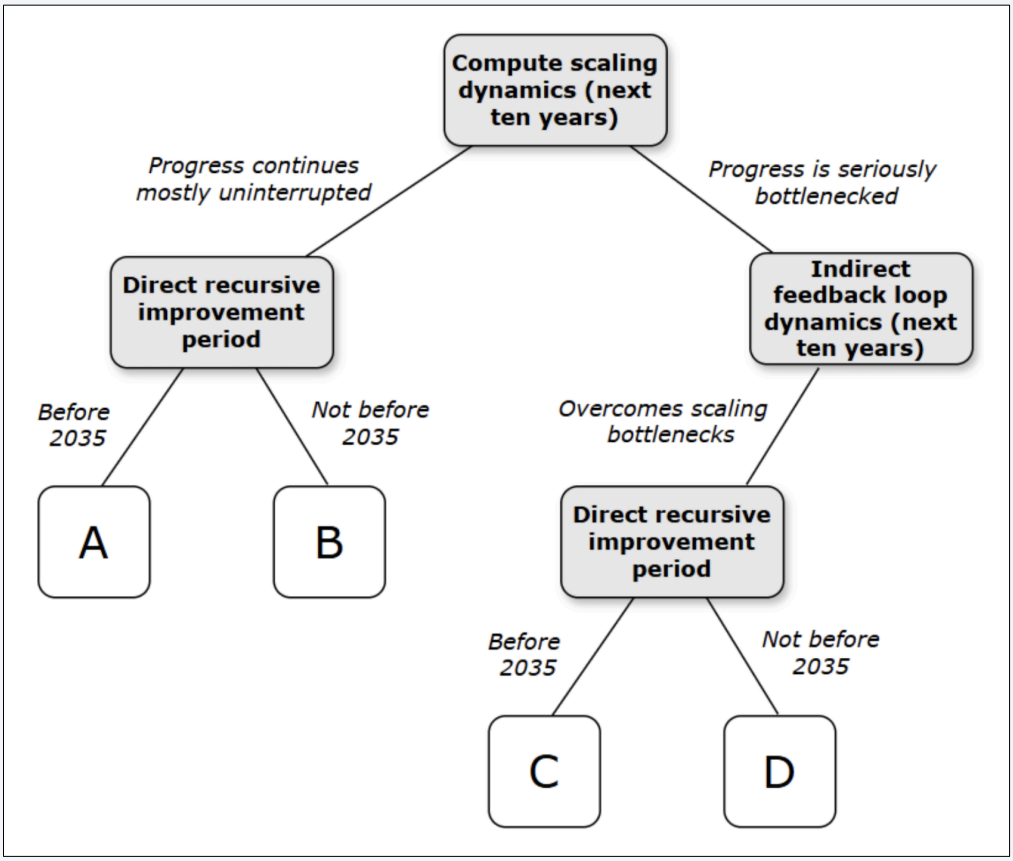
Example tree fragments:



The tree fragment on the right hand side here goes to a degree of granularity which is incoherent in the context of short (i.e. pre-2035) timelines.

Secondly, not every future is meaningful in the context of this report. For example, suppose that compute scaling is seriously bottlenecked over the next ten years, but indirect feedback loops gain sufficient traction to overcome those bottlenecks, restoring the previous rates of progress under compute scaling. In this case, it doesn't add much value to then consider whether a period of DRI also begins before 2035; the two possible outcomes of this additional question would very closely resemble the two corresponding outcomes of the pathway on which compute scaling had continued uninterrupted in the first place.

Example tree fragment:



The outcomes {A,B} here appear to be very similar to {C,D} in terms of their likely trajectories of capabilities growth, as well as the overall plausibility of their resulting in a short timeline. Of course, these two sets of outcomes aren't completely identical: for example, under {C,D}, we might see some additional delays to progress that are not present under {A,B}, corresponding to the time taken for indirect feedback loops to gain enough traction to overcome bottlenecks. But within the wider context of this chapter and the breadth of possible parameter value combinations, the differences here seem minor.

To avoid unnecessary complexity, I limit the selection of parameter 'prompts' in the tree (which are represented as nodes) to those which either yield a *coherent* and *meaningfully new* short timeline scenario, or otherwise result in a 'dead end' (due to implausibility of a short timeline under the given assumptions). This results in a refined set of five short TAI timeline scenarios and two dead ends.

The decision process, in full

Below, I explicate the decision process which defines the structure of the tree.

- → **START: Compute scaling dynamics (next ten years)**
  - ↳ If compute scaling *continues uninterrupted*, **GO TO: Direct recursive improvement period**
    - ↳ If DRI also begins *before 2035*, **END: Scenario 5**
    - ↳ If DRI does *not* begin *before 2035*, **END: Scenario 1**
  - ↳ If compute scaling *is seriously bottlenecked*, **GO TO: Indirect feedback loop dynamics (next ten years)**
    - ↳ If indirect feedback loops gain enough traction to *overcome scaling bottlenecks*, **END: Scenario 2**
    - ↳ If indirect feedback loops *do not overcome scaling bottlenecks*, **GO TO: Direct recursive improvement period**
      - ↳ If DRI begins *before 2035*, **GO TO: Direct recursive improvement strength**
        - ↳ If DRI *cannot sustain existing progress*, **DEAD END**
        - ↳ If DRI *sustains existing progress*, **END: Scenario 3**
        - ↳ If DRI *accelerates progress*, **END: Scenario 4**
      - ↳ If DRI does *not* begin *before 2035*, **DEAD END**