Pathways to Short Transformative AI Timelines

Zershaaneh Qureshi

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.

Correspondence: zqureshi@s-rsa.com

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'. 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

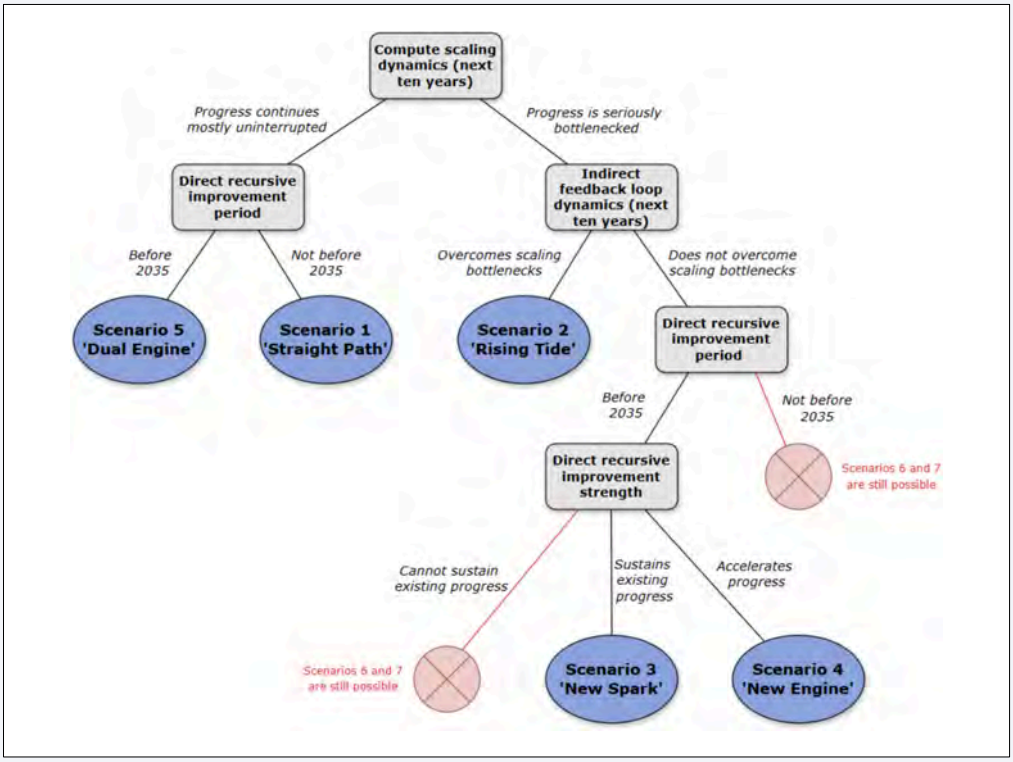


Figure 3.1: The scenario tree. A full-sized version of this figure can also be found in Appendix A.

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]

Direct recursive improvement (DRI): positive feedback loops which are mediated directly by AI systems. *Indirect recursive improvement (IRI):* positive feedback loops that are not mediated directly by AI, such as 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).

[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.

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.

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.

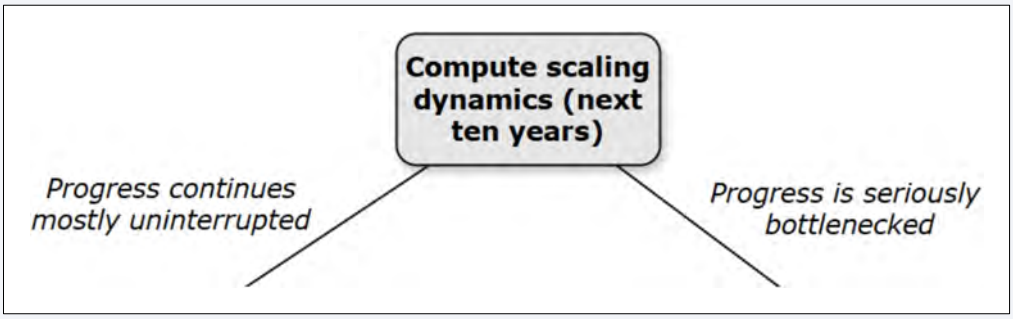
The key parameters, and the values they can take, are outlined below. Screenshots of the relevant nodes of the scenario tree are also included alongside these descriptions.

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*.



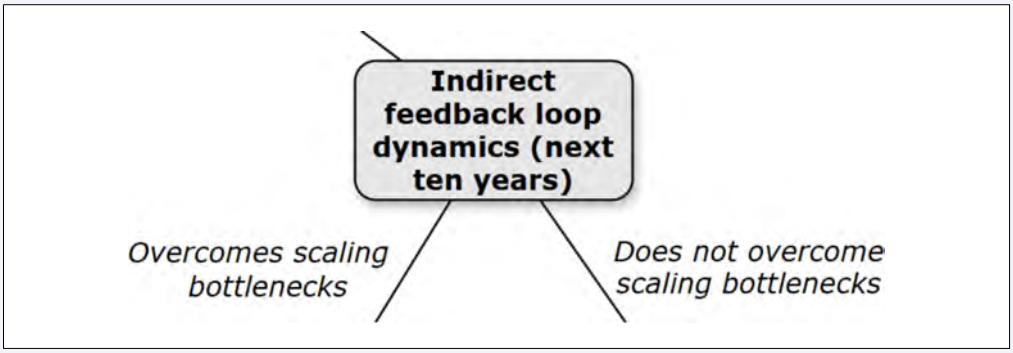
One important determinant of the world we are in is the fate of the current trajectory of capabilities progress, driven by compute scaling. Over the next

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*.



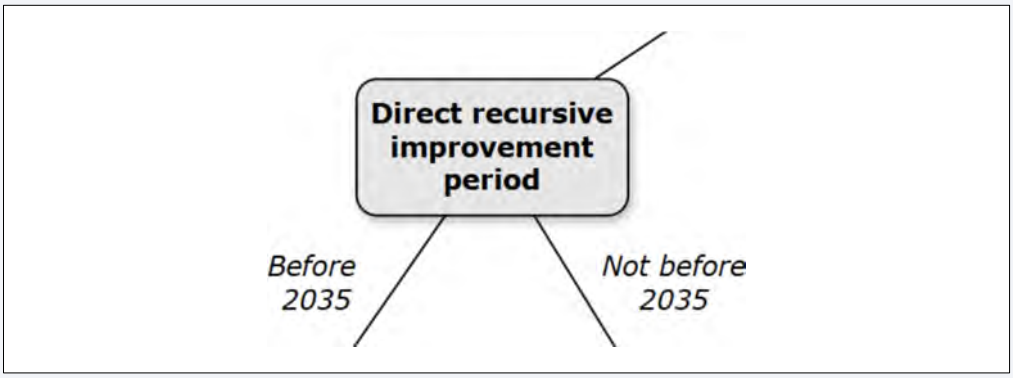
If the compute scaling pathway does get 'seriously bottlenecked' over the next ten years, it is relevant to ask whether 'indirect' feedback loops (whether economic, scientific, or political) will eventually gain enough traction to *overcome* those bottlenecks to compute scaling.

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*.

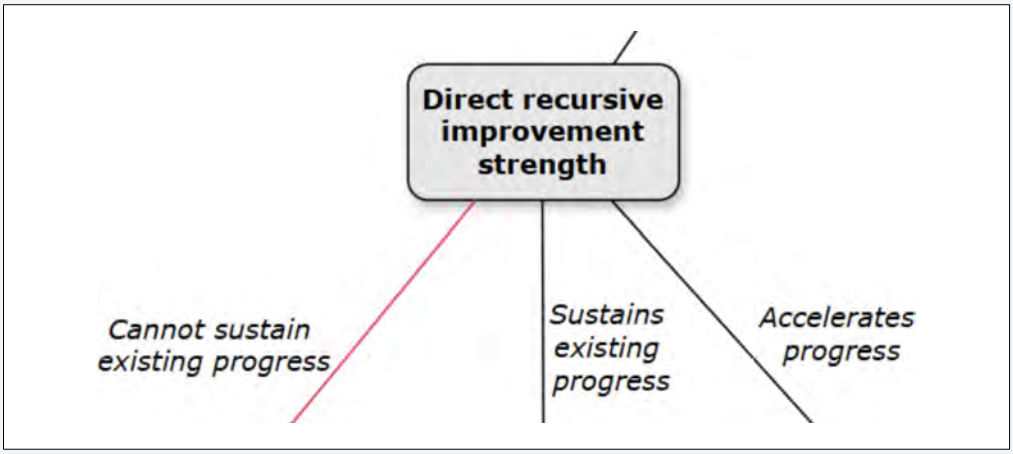


Unlike indirect feedback loops, which are already supporting AI progress to some degree, the emergence of direct feedback loops would mark a meaningful *new* development in the field that could define the next era of capabilities progress.

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.*

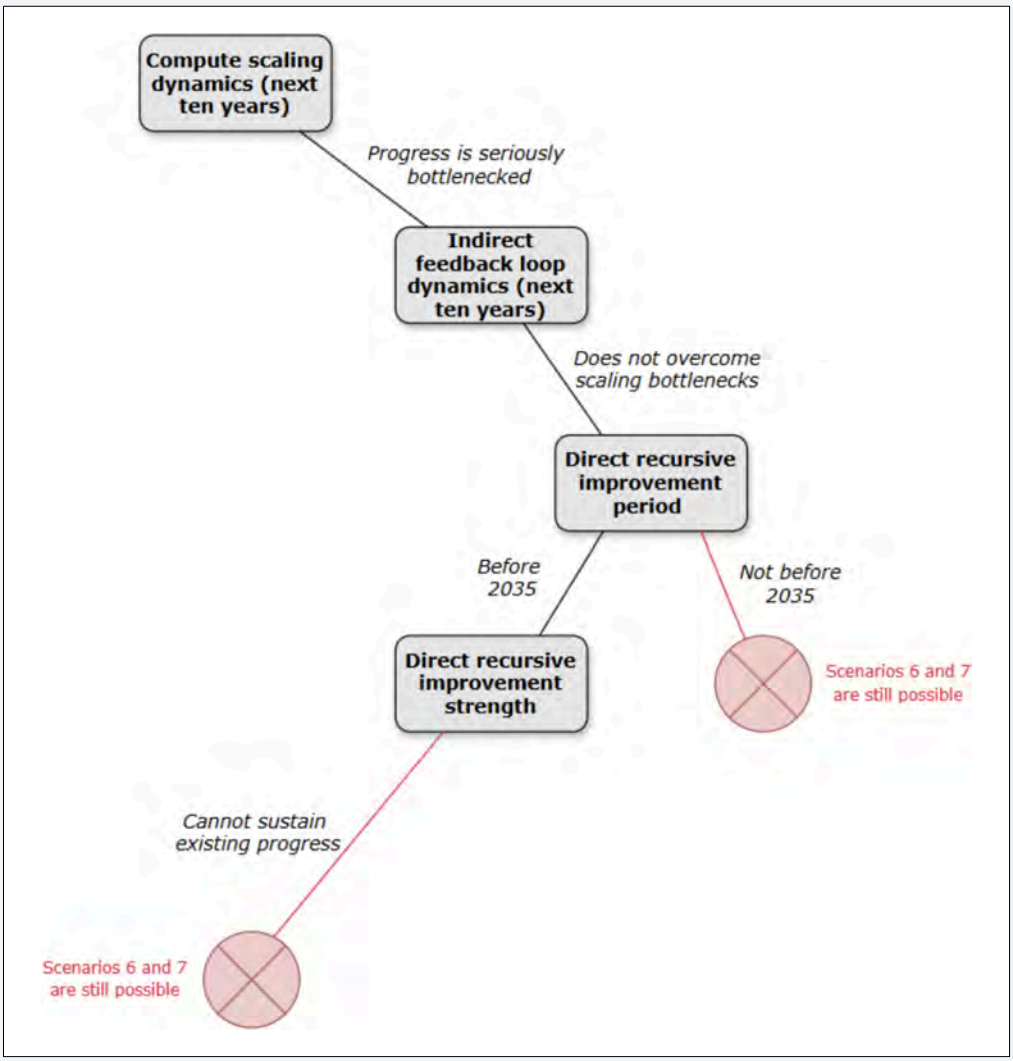
As in Scenario 1, compute scaling with the current paradigm continues to yield results and does not become seriously bottlenecked on anything in the next ten years. There are problems to solve along the way, but there are quick fixes available. Direct recursive improvement also kicks in within the next 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 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.