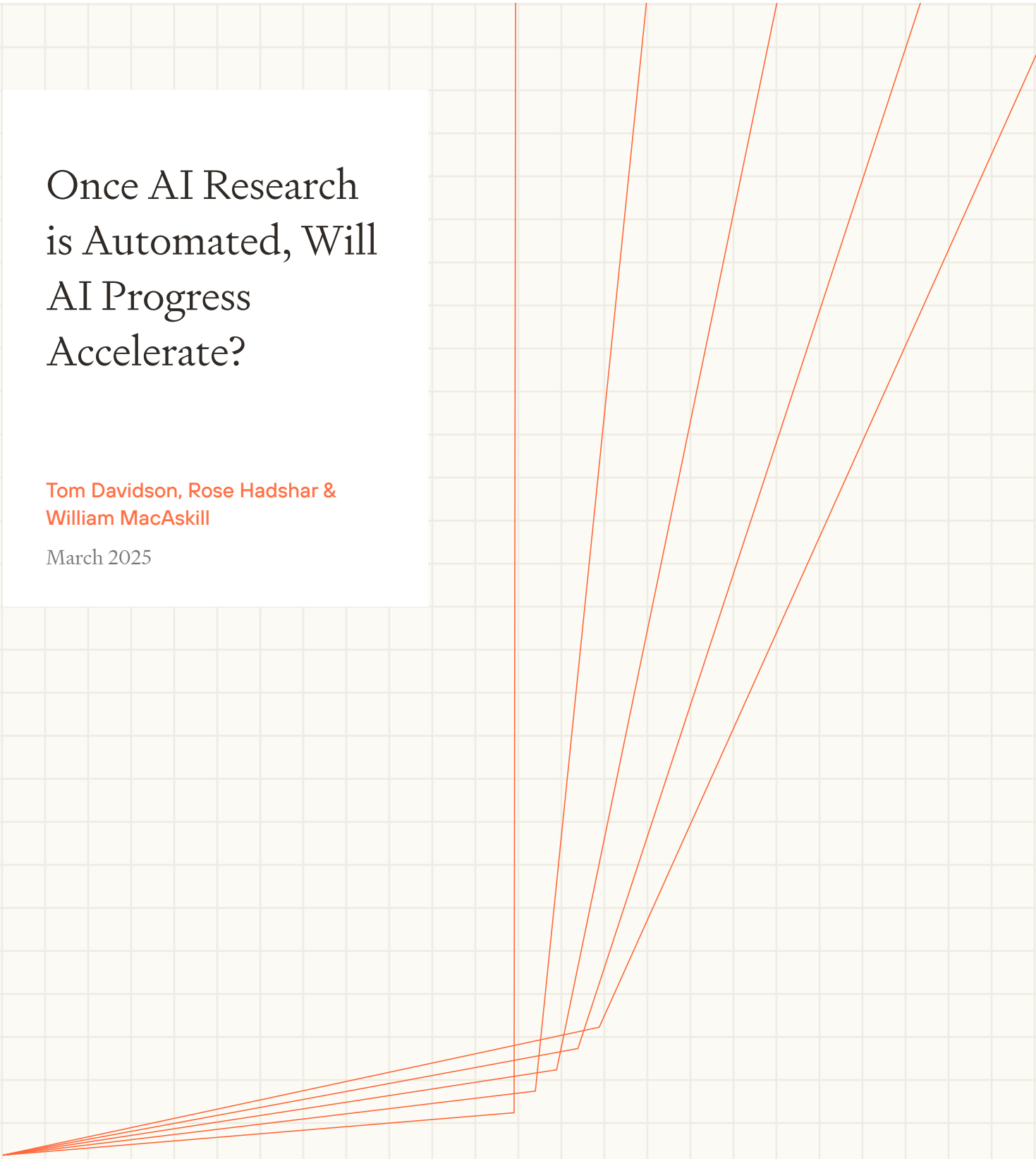
Forethought

Once AI Research is Automated, Will AI Progress Accelerate?

Tom Davidson, Rose Hadshar & William MacAskill

March 2025



Contents

Once AI Research is Automated, Will AI Progress Accelerate?

Tom Davidson, Rose Hadshar & William MacAskill

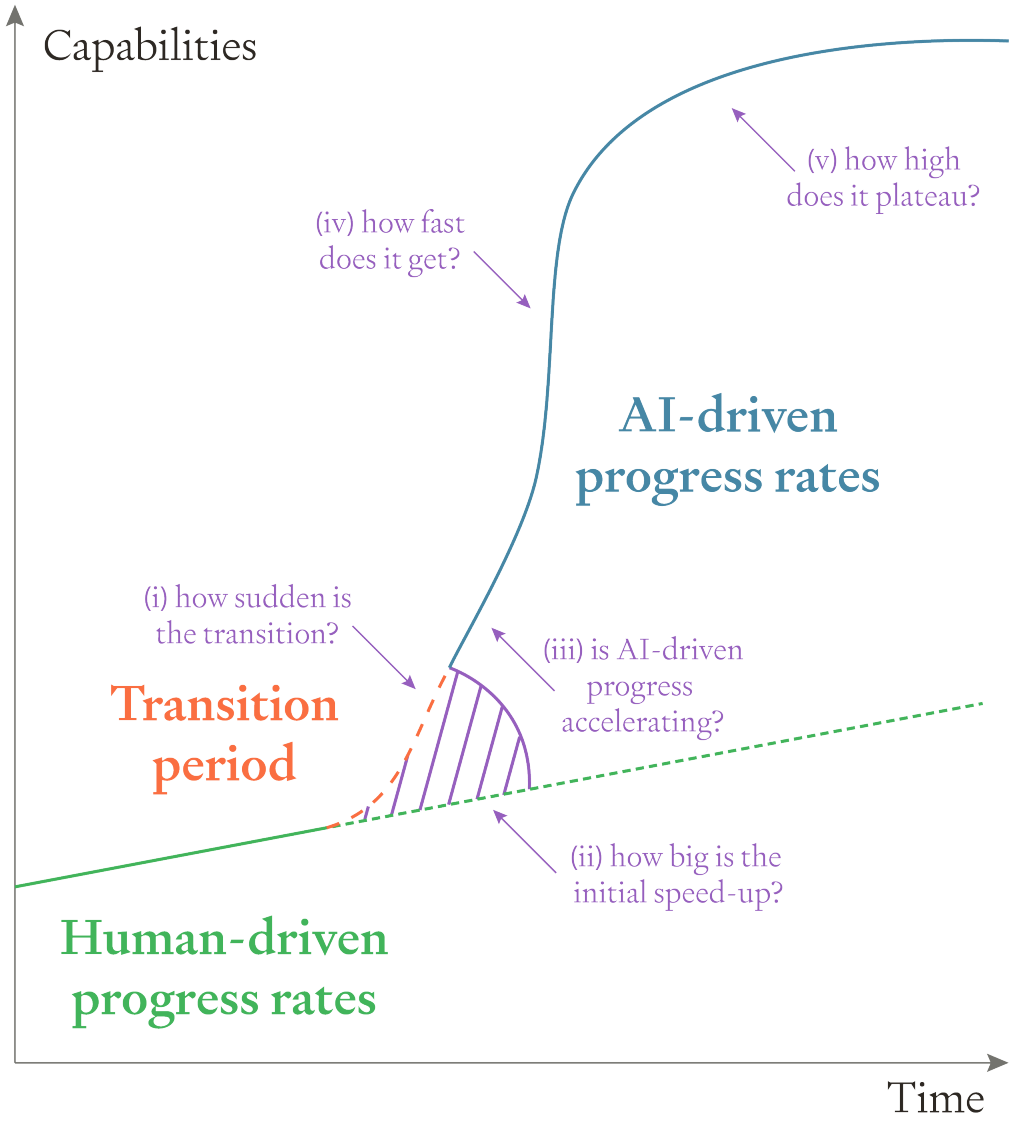
Tom did the original thinking; Will and then Rose helped with later thinking and writing.

For correspondence, please email contact@forethought.org.

This is a rough research note – we're sharing it for feedback and to spark discussion. We're less confident in its methods and conclusions.

Today, AI progress is driven by humans, who write software and design and produce AI chips. But in the future, more and more human work will become automated. Eventually, AI progress will be driven by AI, not humans.

We can imagine a graph which plots capabilities over time, and shows the transition to AI-driven progress as well as the era of AI-driven progress:



In the graph AI progress is initially driven by human efforts (green line). Then over time AI increasingly automates the work for improving AI (orange line). Eventually, AI does almost all the work for improving AI (blue line).

The graph raises several different factors which are relevant to understanding the landscape of the transition to AI-driven progress:

i. **Suddenness :** how many months or years does it take to transition from human-driven progress to AI-driven progress?

ii. **Initial speed-up :** how much faster is the *initial* period of AI-driven progress, compared to the final period of human-driven progress?

iii. **Acceleration:** after the transition, will AI-driven progress *accelerate* over time?

iv. **Peak growth rate :** how fast could AI progress eventually become?

v. **Plateau :** how far can AI progress before hitting effective physical limits?

In this post, we're going to discuss (iii): once AI has fully automated the work to improve AI, will AI progress accelerate over time?

This is an important question, because if progress accelerates, we could end up with much larger and faster increases in AI capabilities.

First, we explain the conditions under which progress will accelerate in general terms.

Then we evaluate whether these conditions hold for three separate feedback loops via which AI will improve AI:

- A *software* feedback loop, where AI develops better software. Software includes AI training algorithms, post-training enhancements, ways to leverage runtime compute (like o1), synthetic data, and any other non-compute improvements.
- A *chip technology* feedback loop, where AI designs better computer chips. Chip technology includes all the cognitive research and design work done by NVIDIA, TSMC, ASML, and other semiconductor companies.
- A *chip production* feedback loop, where AI and robots build more computer chips.

We argue that, putting aside any human actions to slow AI progress:

- **The software feedback loop** *by itself* might well be enough to sustain accelerating progress (~50%).
- **The chip technology feedback loop** *by itself* is probably enough to sustain accelerating progress (~65%).
- **The chip production feedback loop** *by itself* is very likely enough to sustain accelerating progress (~80%).

We conclude by arguing that overall AI progress is even more likely to accelerate , as it can involve all three feedback loops.

(This piece gives qualitative explanations only, and is based on forthcoming mathematical work by Basil Halperin and Thomas Houlden.)

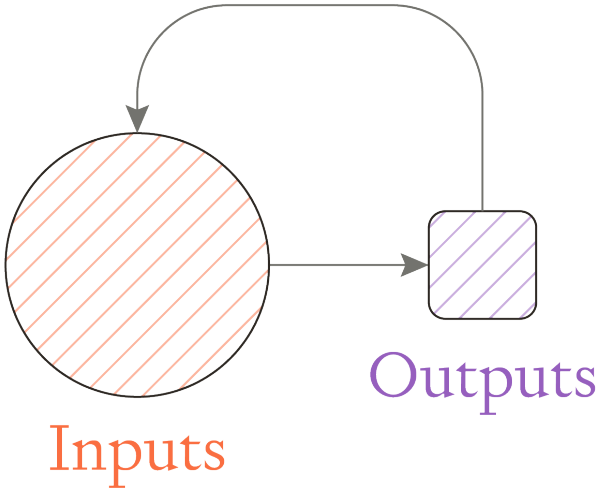
Under what conditions does progress accelerate?

AI improving AI is not the only example of a positive feedback loop. Before discussing AI in particular, we'll consider the general conditions under which a positive feedback loop leads to accelerating growth.

Let's imagine a simple system that takes inputs and produces outputs. For example, we can think of the global economy taking inputs of human labour and physical capital, and outputting goods and services.

If the outputs can be used to create more inputs, then there's a positive feedback loop. (For example, goods and services are used to build more physical capital.)

Feedback loop



The existence of a positive feedback loop, however, doesn't mean that you get accelerating progress. For acceleration, each doubling of inputs must take less time than the previous doubling - even though the volume of inputs is twice as large.[1]

The key issue is: **when you double the (cumulative) inputs,[2] how many times do the outputs double?**

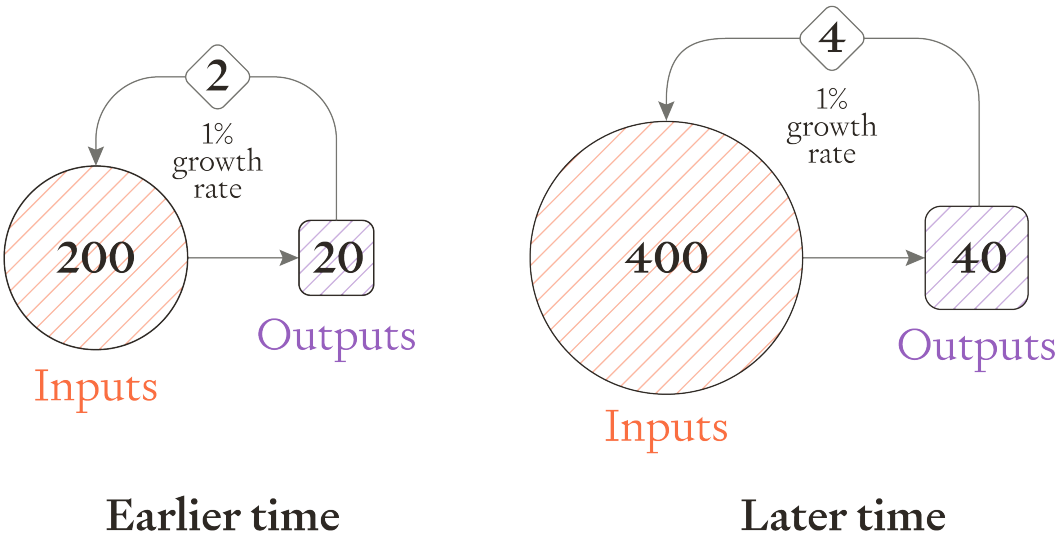
If outputs double exactly once, then you have constant exponential growth – no acceleration.

---

1 This assumes that there is some power law relationship between inputs and outputs.

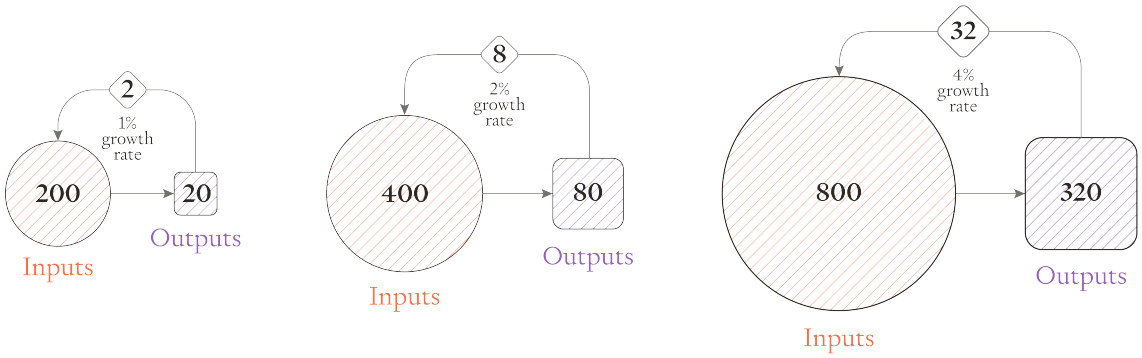
2 We're interested in cumulative inputs because current levels of technology (the relevant meaning of 'output') build on the total effort that's gone into the problem so far. In areas of technology, if you double your research workforce overnight, the level of technology won't immediately double. So we look at how many times output doubles when *cumulative inputs* , that is, inputs added up across time, double. On this definition, having 200 people work for a year constitutes twice as much cumulative input as having 100 people work for a year.

To get a feedback loop, the inputs also need to be accumulable: if you get more outputs you can convert them into more inputs.



At the earlier time, 200 units of input produced 20 units of output. 2 of these units of output are reinvested as inputs (a 10% reinvestment rate), so inputs have a growth rate of 1%. To grow at the same rate once inputs have doubled, you need to produce double the outputs (assuming a constant reinvestment rate).

And if outputs *more than* double, you get acceleration: at the later time you have twice as many inputs (so growing them by 1% takes twice as much effort) but you have *more than* twice as many outputs.



When inputs double, outputs more than double. As a result, the growth rate increases.

In some growth models , the number of times outputs double when cumulative inputs double is called *r* .[3] If r>1, outputs double more than once and you get accelerating growth. If r is below one, outputs less than double when you double cumulative inputs, so growth doesn't accelerate.

For example, over the long-run course of history, doubling the number of humans has (on average) led to *more* than a doubling of output. (With twice as many humans, you can have twice as many people farming. But those people can *also* be coming up with new ideas for better farming methods, so that the amount of food produced - and humans that can be supported in future

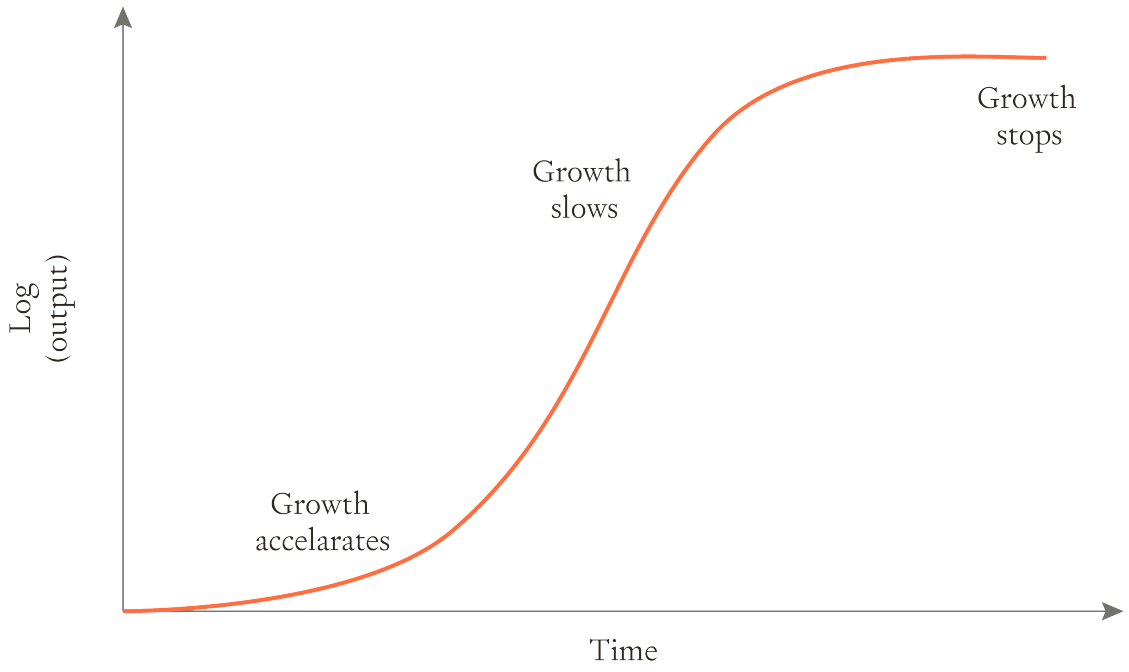
3 Technically speaking, r is defined as follows: if your inputs grow exponentially at g%, your outputs will grow exponentially at g*r%.

generations - more than doubles.) And this meant that (on average) the next doubling of humans happened more quickly, as r was greater than 1.

Year	Human population	Years to double
0 CE	~200m	1100
1100	~400m	660
1760	~800m	140
1900	~1.6bn	60
1960	~3.2bn	40
2000	~6.4bn	-

Each doubling of the human population has taken less time, even though each increase was twice as large.

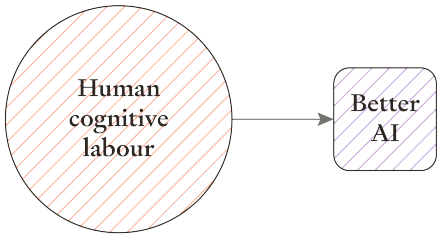
In domains where doubling inputs *does* lead to more than a doubling of outputs, the returns to doubling inputs will eventually start to diminish more steeply, r will drop below 1, and growth will stop accelerating — for example once the system approaches some sort of physical limit. So, in practice, accelerating growth is always just one part of an s-curve:



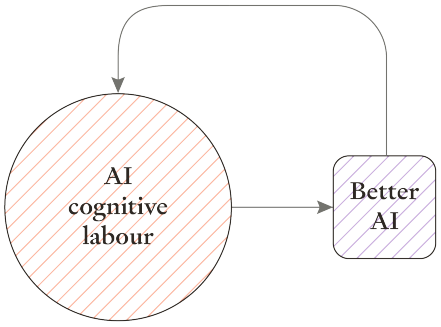
Will AI progress accelerate?

For AI progress, we're thinking about the same accelerating growth dynamics, but with cognitive rather than economic inputs and outputs.

Currently, humans do almost all the cognitive labour for improving AI, and so there isn't a strong feedback loop of better AI → more cognitive labour for improving AI → better AI. But if AI replaced humans at the task of improving AI, the resulting feedback loop could lead to accelerating AI progress.



Today, better AI doesn't increase the amount of human cognitive labour (much).



In future, better AI will directly increase the amount of AI cognitive labour.

We can define "accelerating AI progress" as "each increment of capability advancement (e.g. GPT-2 → GPT-3 or GPT-3 → GPT-4) happens more quickly than the last".

As in the general case of accelerating growth, the key question is how many times the outputs double for each doubling of AI inputs:[4] i.e. what is r? If r>1, AI progress accelerates. If r<1, it doesn't.

We can get empirical evidence on r for each of the three feedback loops, and use this to guess the likelihood of acceleration.

We argue that:

- For the **software feedback loop** , r is around 1, and possibly slightly higher.
- For the **chip technology feedback loop** , r is comfortably above 1.
- For the **chip production feedback loop** , r is robustly above 1.

Note that in this analysis we focus on *technological* likelihood only, and ignore any human bottlenecks like regulation or cultural constraints which might limit the pace of acceleration. The all-things-considered likelihood of acceleration will be lower.

4 For the software and chip technology feedback loops, AI inputs are cognitive only. For the chip production feedback loop, which includes robots, AI inputs include physical inputs too.

Software feedback loop

The question we care about is, when we double the human researcher population (i.e. double cognitive inputs), how many times does software double?[5] If the answer is >1, then progress is accelerating.

The most relevant data is on efficiency improvements:

- **Training efficiency of ImageNet.** The most relevant estimate is Epoch's analysis of Imagenet training efficiency, which estimates r = ~1.4.
- **Other data on efficiency gains.** Epoch also analysed runtime efficiency in other domains:
  - Computer chess: r = 0.8
  - Problem solvers: r = 3.5
  - Linear programming: r = 1.1

But we don't just care about efficiency gains – getting the same capabilities for less compute – we also care about improving capabilities – the ability to do tasks we previously couldn't do.[6] This increases the estimate significantly. After considering this and downwards adjustment factors like limited compute for running experiments, Davidson and Houlden (2025) estimate r = 1.2, with a range of 0.4 to 3.6.

Eth and Davidson review the evidence for software improvements, and argue that, after accounting for compute-experiment bottlenecks, a reasonable range for r today is 0.5 - 2.

We think it is ~50% likely that the software feedback loop drives accelerating progress, absent human bottlenecks**.**

Chip technology feedback loop

Again we can ask the question, when we double the human researcher population, how many times does chip technology double?

- Evidence from Davidson's compute-centric framework report suggests that recently r = ~5 when considering all inputs to hardware R&D.[7] If we assume that human cognitive labour is 60% of all hardware R&D inputs, then r = ~3.
- As with software, doubling chip technology is actually better than doubling the human researcher population (as you don't just double the quality of your hardware, you consequently train a model with higher capabilities). This might approximately double r, to ~6.[8]

---

5 What does it mean for software to "double"? We can think about improvements in training efficiency leading to capability gains. To incorporate such improvements precisely, we can translate them into an "equivalent" runtime efficiency gain. If some capabilities improvement increases our AI's ability to make research progress by the same amount as doubling efficiency, we'll say that capabilities improvement "doubled" software. In other words, all software improvements will be measured in the same units, where a "doubling of software" corresponds to any improvement that's "as useful" as doubling runtime efficiency.

6 If we double the quality of our software, our AI models will be more efficient by some factor, but they will also be more capable.

7 This estimates r for FLOP/$, which includes both improving the energy efficiency of chips (FLOP/Joule) and making chips cheaper to make and run ($/chip and $/Joule).

- We should adjust downwards as r will be falling over time (as the low-hanging fruit is picked), perhaps to r = ~4.
- Experts argue that we are approaching fundamental limits to the current paradigm.[9] We can make a further downward adjustment in deference to these experts, but r still seems likely to be comfortably above 1.

We think it is likely (very roughly ~65%) that the chip technology feedback loop drives accelerating progress , absent human bottlenecks.

Chip production feedback loop

Because software and (our definition of) chip technology are virtual, better AI straightforwardly leads to better software and chip technology. But chip production also involves physical labour, so there is a key additional step in the feedback loop: robots (and other machines and infrastructure) which can fully automate the process of building chip factories.

With robots, a doubling of inputs would straightforwardly result in a doubling of outputs. That's because if you have twice as many robots, they can *build* twice as many. If robots also improve robot technology at all, which seems likely,[10] then output would more than double for each doubling of input.

On the other hand, doubling inputs might *less* than double outputs if scarce natural resources take more and more work to extract, but historically when raw materials have become scarce this has been more than compensated for by innovation.[11]

(This is simply the argument given for accelerating economic growth – that once you fully automate labour the economy displays increasing returns to scale. See Erdil and Besiroglu (2023) and Davidson (2021) .)

We think it is highly likely (very roughly ~80%) that the chip production feedback loop drives accelerating progress, absent human bottlenecks.

Will intelligence explosions accelerate?

What does this mean for whether an intelligence explosion would accelerate? Elsewhere , we argue that these feedback loops could lead to three particularly plausible kinds of intelligence explosion (IE):

- A *software* IE, where AI-driven software improvements alone are sufficient for rapid and accelerating AI progress.

---

8 Davidson estimates that doubling training size doubles output-per-FLOP at runtime between once and twice. If we are conservative and take the lower number, then doubling hardware 1) doubles your runtime FLOP and 2) doubles your output-per-FLOP at runtime, i.e. it doubles your total output twice.

9 For instance, here and here .

10 More robots can build more chips which can be used to run more AI models to work on improving robot technology. So having more robots seems likely to lead to improved robot technology.

11 This will eventually cease once all the materials on earth or the solar system are used up, but there is a lot of room for acceleration before that point.

- An *AI-technology* IE, where AI-driven improvements in both software and chip technology are needed, but AI-driven improvements in chip production are not.
- A *full-stack* IE, where AI-driven improvements in all of software, chip technology and chip production are needed.

When thinking about how these feedback loops combine to create intelligence explosions, we can simply add up the r values for each feedback loop involved.[12] If the total r is greater than 1, then progress will accelerate.

This means that it's more likely that the intelligence explosions accelerate, than that the feedback loops in isolation do.

Another important point here is that there's some correlation between the r values:

- If 'extra compute' doesn't translate to extra intelligence, that weakens both the chip technology and chip production feedback loops.
- If extra intelligence doesn't actually help with research (e.g. because humans are already close to hitting non-cognitive bottlenecks in many domains), that weakens all of the feedback loops.

So lower r values for one feedback loop is correlated with lower r in others. This makes the extremes of very fast acceleration or no acceleration at all more plausible.

Overall, we think that:

- **A software IE might well accelerate over time,** because the software feedback loop *by itself* might well be enough to sustain accelerating progress (~50%).
- **An AI-technology IE would likely accelerate.** The chip technology feedback loop *by itself* is probably enough to sustain accelerating progress (~65%). This means the *combination* of the software and chip technology feedback loops are likely jointly strong enough to drive accelerating progress (~75%).
- **A full-stack IE is highly likely to accelerate.** It's likely that the chip production feedback loop *by itself* can sustain accelerating progress (~80%). So in combination with the other feedback loops it is highly likely that an industrial IE would accelerate (~90%).

Thanks to Max Dalton, Oscar Delaney, Daniel Eth, Fin Moorhouse and the governing explosive growth seminar group for helpful comments.

12 See Houlden, Thomas, presenter. Is Automating AI Research Enough for a Singularity? By Tom Davidson, Basil Halperin, Thomas Houlden, and Anton Korinek. 15th Oxford Workshop on Global Priorities, December 2024.