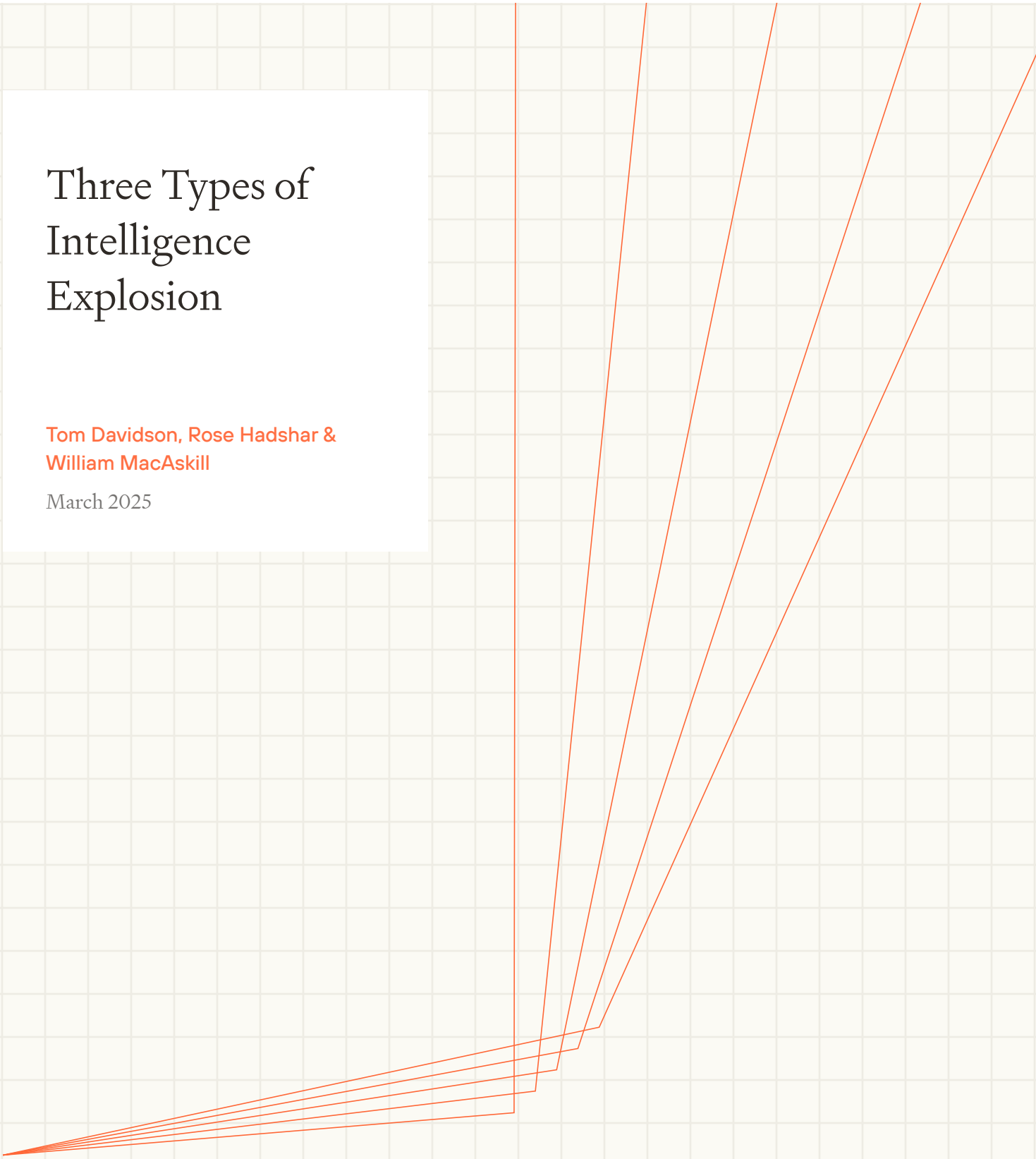
Forethought

Three Types of Intelligence Explosion

Tom Davidson, Rose Hadshar & William MacAskill

March 2025



Contents

Three Types of Intelligence Explosion

Summary 3

Three types of feedback loop 5

Time lags in each feedback loop 6
Order of the feedback loops 9
Three kinds of intelligence explosion 9

Analysis of these feedback loops 12

Will AI progress accelerate over time? 12
How far could AI progress before hitting effective physical limits? 15
The landscape of the intelligence explosion 17

Strategic implications for the distribution of power 20

Appendix: Are there really just three feedback loops? 22

Appendix: How fast could AI progress eventually become? 23

Tom Davidson, Rose Hadshar & William MacAskill

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

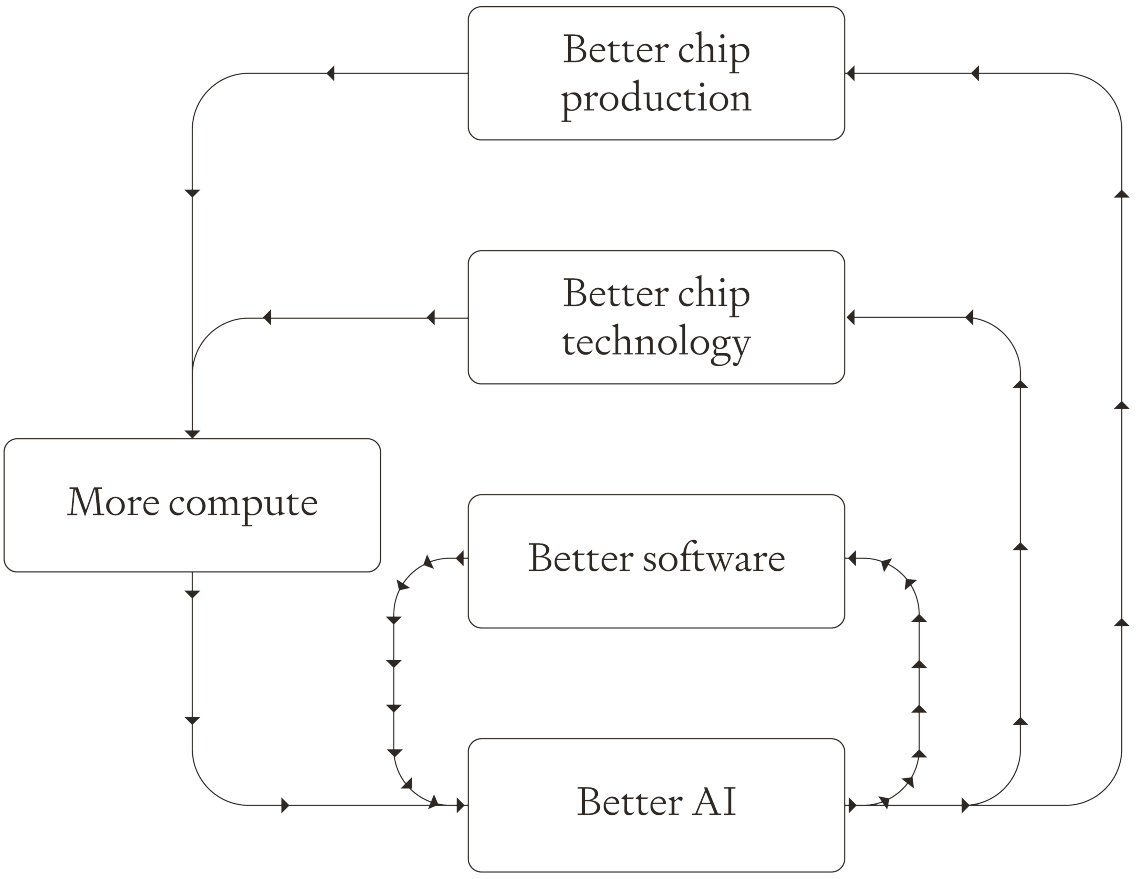
For correspondence, please email contact@forethought.org.

Summary

Once AI systems can themselves design and build even more capable AI systems, progress in AI might accelerate, leading to a rapid increase in AI capabilities. This is known as an *intelligence explosion* ("IE").

The classic IE scenario involves a feedback loop in AI software, with AI designing better software that enables more capable AI that designs even better software, and so on. But there are many parts of AI development which could lead to a positive feedback loop. We identify:

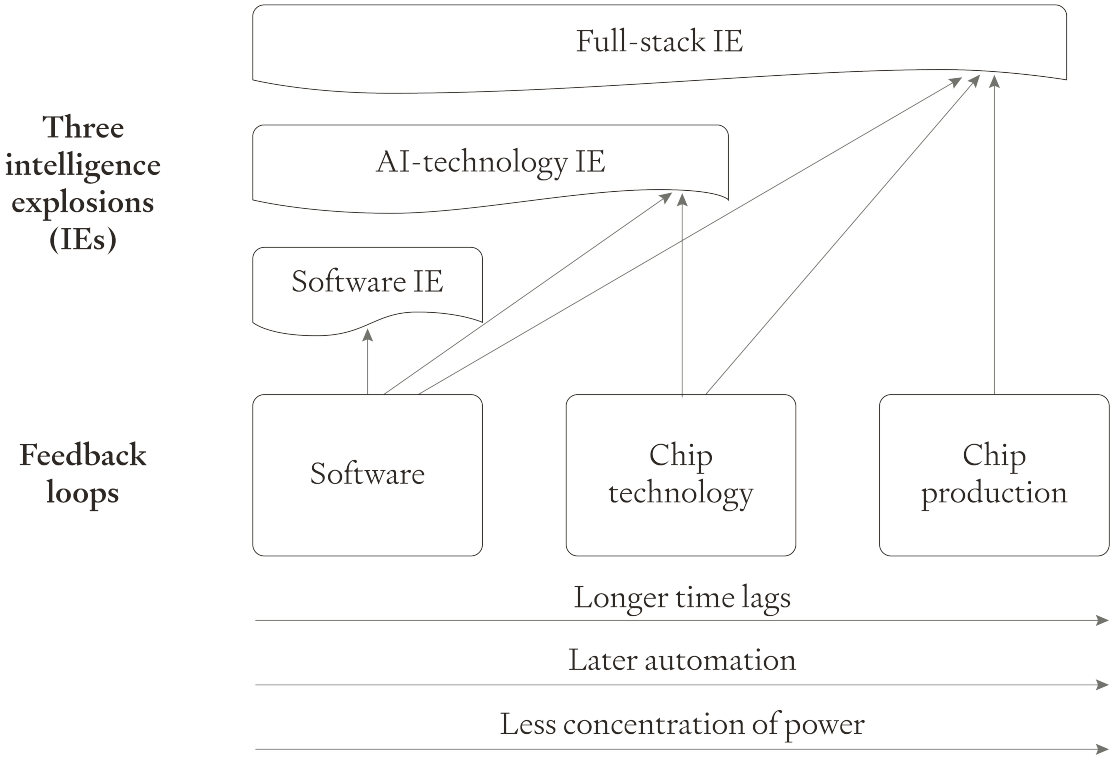
- A *software* feedback loop, where AI develops better software. Software includes AI training algorithms, post-training enhancements, ways to leverage runtime compute (like o3), synthetic data, and any other non-compute improvements.
- A *chip technology* feedback loop, where AI designs better computer 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.



The three feedback loops.

The software loop will likely be automated first and it has the shortest time lags (training new AI models), and the chip production loop will likely be automated last and has the longest time lags (building new fabs). These feedback loops could drive three different types of IE:

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



The three intelligence explosions.

Crucially, even if the software feedback loop is not powerful enough to drive a software IE, we could still see an AI-technology or full-stack IE.

An IE is more likely if progress accelerates after full automation. We think, based on empirical evidence about diminishing returns, that the software and AI-technology IEs are more likely to accelerate than not, and that a full-stack IE is very likely to accelerate eventually.

An IE will be bigger and faster if effective physical limits are further away. We estimate that before hitting limits, the software feedback loop could increase effective compute by ~13 orders of magnitude ("OOMs"), the chip technology loop by a further ~6 OOMs, and the chip production feedback loop could increase effective compute by a further ~5 OOMs (and by another 9 OOMs if we capture all the sun's energy from space).

If the recent relationship between increasing effective compute and increasing capabilities continues to hold, this would be equivalent to ~4 "GPT-sized" jumps in capabilities from software

(i.e. 4 jumps as large as the jump from GPT-2 to GPT-3, or GPT-3 to GPT-4), a further ~2 GPTs from chip technology, and a further ~2-5 GPTs from chip production.[1]

These IEs differ in their strategic implications. A software IE would be most likely to occur first in the US, with power strongly concentrated in the hands of the owners of AI chips and algorithms. An AI-technology IE would most likely involve the US and some other countries in the semiconductor supply chain like Taiwan, South Korea, Japan, and the Netherlands, with power more broadly distributed among the owners of AI algorithms, AI chips and the semiconductor supply chain. Compared to the other two IEs, a full-stack IE may be more likely to heavily involve countries like China and the Gulf states, which have a strong industrial base and a more permissive regulatory environment. A full-stack IE would also distribute power more broadly across the industrial base.

Three types of feedback loop

In 1965, mathematician I. J. Good introduced the idea of an intelligence explosion:[2]

"Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make."

The core idea is that once AI systems can themselves design and build even more capable AI systems, then there would be a feedback loop where AI creates better AI which creates even better AI... If this feedback loop causes AI progress to become very fast, we could call the result an *intelligence explosion* (IE).[3] This post explores how the intelligence explosion might be structured.

The classic IE scenario involves a feedback loop in AI software, with AI designing better software that enables more capable AI that designs better software, and so on. But there are actually many parts of AI development which could lead to a positive feedback loop. We can distinguish *three*[4] important feedback loops that could drive an intelligence explosion:

1. A **software feedback loop** , where AI improves AI algorithms, data, post-training enhancements, and other software techniques. The central example of this feedback loop is fully automating the research and engineering work at frontier AI development labs. Here, AI systems improve algorithms which are used to develop better AI systems, which further improve algorithms, and so on.

---

1 In recent years, increasing the effective compute for AI development by 1000X has been sufficient to improve capabilities by 1 GPT (so 1 GPT is 3 OOMs of improvement). See workings here .

2 Good, I.J. (1966) 'Speculations Concerning the First Ultraintelligent Machine', in F.L. Alt and M. Rubinoff (eds) Advances in Computers. Elsevier, pp. 31–88. Available at: https://doi.org/10.1016/S0065-2458(08)60418-0 .

3 Defining exactly how fast AI progress needs to be to qualify as an IE is ultimately fairly arbitrary. In this post we don't make use of a specific definition.

4 The choice of *three* feedback loops, rather than fewer or more, is natural but somewhat arbitrary. See here for further discussion.

2. A **chip technology feedback loop** , where AI improves the *quality* of AI chips.[5] The central example is automating all *cognitive* work done by the R&D functions of NVIDIA, TSMC, ASML and other semiconductor companies. This R&D work allows for the production of better chips (without increasing the number of factories) – that is, chips with higher FLOP/s and other improved specs. By training (or running inference) on these higher-compute chips, AI capabilities improve. This creates more and better cognitive labour to put towards semiconductor R&D, leading to the production of even more effective chips, and so on.[6]
3. A **chip production feedback loop** , where AI increases the *quantity* of AI chips. The central example is robots fully automating the process of building and running chip factories, including mining the raw materials, transporting them, building the factories, and operating them. These robots build more chip factories and more chips, which are used to train more and better AI systems, which then design better robots to build even more chip factories, and so on.[7]

Time lags in each feedback loop

How suddenly could an intelligence explosion happen - how quickly could we transition from human-driven progress in AI capabilities to AI-driven progress in AI capabilities?[8]

One reason things are unlikely to be extremely sudden is that all the feedback loops have inbuilt time lags.

---

5 Automated chip technology could also increase the *quantity* of AI chips, via process improvements which could enable the same number of fabs to produce more chips than before.

6 The chip technology and chip production feedback loops can improve AI because, even holding AI software fixed, increasing the compute used for AI development and inference improves AI capabilities. Further, increases in inference compute enable the same software to be run more times in parallel; in the context of automated researchers, this would be akin to expanding your researcher pool.

7 The chip technology and chip production feedback loops can improve AI because, even holding AI software fixed, increasing the compute used for AI development and inference improves AI capabilities. Further, increases in inference compute enable the same software to be run more times in parallel; in the context of automated researchers, this would be akin to having more researchers.

8 We can think about this precisely in terms of the elasticity of AI R&D output from humans and AI: how much the marginal pace of progress increases if you increase the population of humans or AIs by 1%. We can say that the transition period begins when AI elasticity is first >1/10 that of human labour, and ends when it's >9/10. The specific start and end points are arbitrary, but for any such choice there's a big range of possibilities for:

- How sudden the transition period could be: it could take anything from months to decades.
- How fast the initial speed-up from AI-driven progress could be: the pace of AI progress could increase by 2X or 100X over the course of the transition.

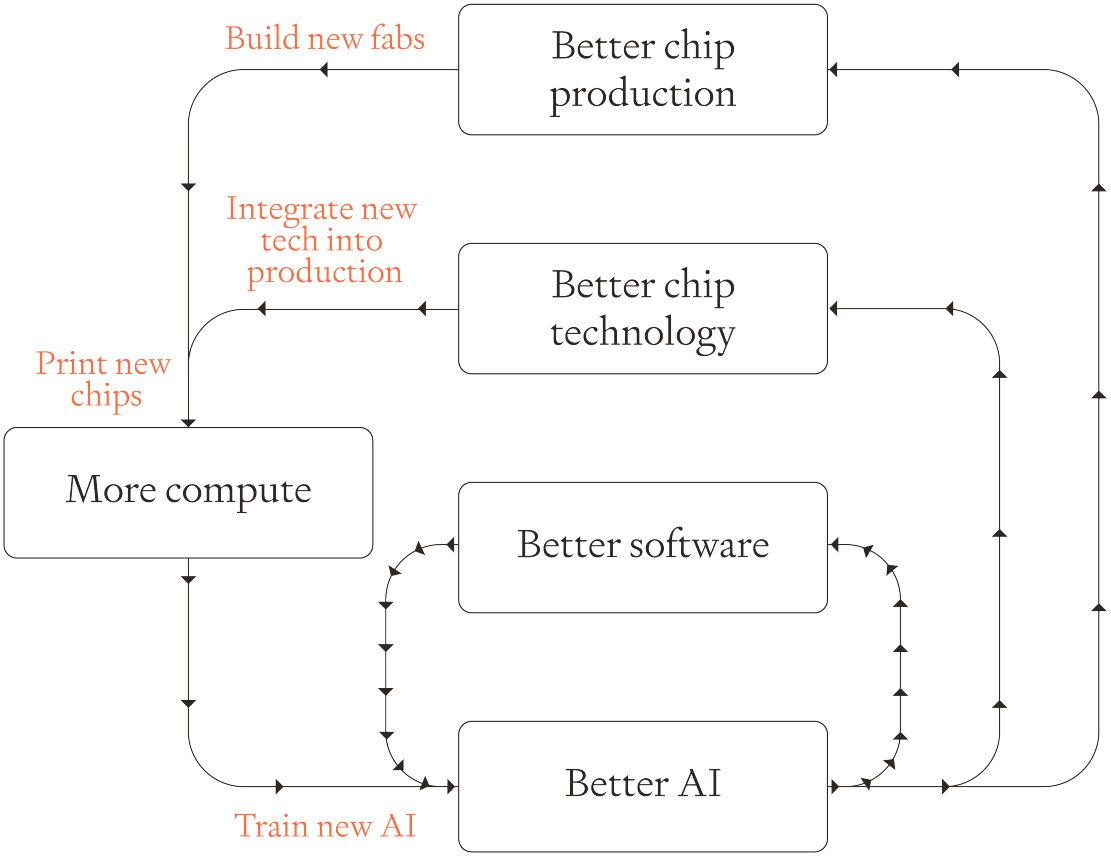
See this essay for further discussion.

Software has the shortest time lags, and chip production probably has the longest:

- It takes ~3 months to train new SOTAn AI, which is the main time lag in the software feedback loop (though post-training enhancements, such as fine-tuning, take *much* less time).
- For the chip technology feedback loop, it's also necessary to integrate new technology into chip factories and print a stock of new chips, which would take many months.[9]
- It takes years to build new fabs, which would be needed for the chip production feedback loop.

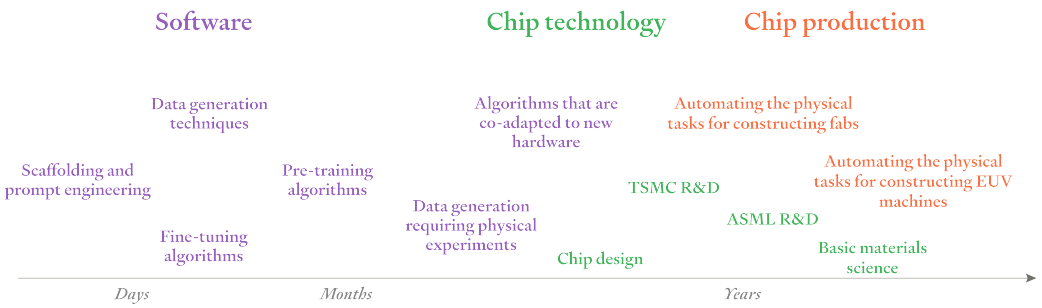
9 New chip designs can begin printing within weeks. But it would take longer to print as many chips as the existing stock. As a rough estimate:

- NVIDIA's AI chip revenues have doubled every ~12 months in recent years.
- If we use this as a proxy for the doubling time of AI chips, then it would take 12 months to print as many chips as have already been made.
- It could take less than 12 months to print the chips for frontier training runs, if the proportion of AI chips spent on frontier training runs is increased.



The three feedback loops, with time lags shown in red. These lags might get shorter during an IE, as AI drives technological progress.

So, all else equal, software improvements have the shortest time lags, followed by chip technology improvements, and increases in chip production have the longest time lags.



Time lag (lag from "work done in this area" to "AI is better")

We could subdivide the three feedback loops further, and visualise the duration of the time lags for each subdivision, as in this diagram.

Order of the feedback loops

In principle an intelligence explosion could contain any combination of these three feedback loops in any order. But we think the software feedback loop is likely to begin first, followed by chip technology, and then finally chip production:

- **Software** is entirely virtual, and so is likely to be automated first. Additionally:
  - AI labs can generate all of the data needed for automation themselves.
  - AI labs have direct access to their own workflows, which makes them easier to automate.
  - Software will initially have shorter feedback loops than chip technology or chip production, so there's more incentive to automate software for actors who want to generate fast AI progress.
- **Chip technology** is also virtual according to our definition, but is likely to be automated after software because:
  - The tasks involved often rely on the specialised knowledge of individual researchers who pass their knowledge from master to apprentice (and who don't work at AI labs), so it's harder to get data to train on.
  - The tasks involved in chip technology are more varied than those involved in software, so it will take longer to automate all of the tasks.
  - It's also harder to generate automated feedback on whether the task is done well, as hardware R&D experiments must be done in the physical world.
- **Chip production** involves very wide-ranging cognitive and physical tasks across every part of the semiconductor supply chain, so is likely to be automated last. Notably, robotics has been a relatively slow area of AI to progress, and this step would involve advanced robotics.

This order also has implications for how an actor trying to accelerate AI progress would prioritise their efforts. First they would prioritise the sources of AI improvement with the shortest time lags (software, especially post-training enhancements). As these run out, they would shift their efforts to the alternatives with the next-shortest time lags (e.g. hardware design by NVIDIA that doesn't require refitting chip factories). They would also try to reduce the time lags (e.g. inventing new techniques for building chip factories more quickly).

Three kinds of intelligence explosion

If these feedback loops are strong enough, they could lead to an intelligence explosion, where AI capabilities progress very rapidly.[10]

Again, in principle an intelligence explosion could contain any combination of these three feedback loops in any order. But in practice:

10 We are not going to define a threshold for an intelligence explosion in this piece; as we use the term, the important thing is that a positive feedback loop where AI improves leads to fast progress in AI capabilities.

- We've argued above that the likely order is software > chip technology > chip production, both in terms of which feedback loop will have the smallest time lags and in terms of which will be automated first.
- The feedback loops are likely to stack cumulatively, with earlier feedback loops still operating when later feedback loops begin.[11]

This suggests that there are three kinds of intelligence explosion which are particularly plausible:

- The software feedback loop alone could lead to a **software IE,** which would have the greatest potential to happen suddenly.[12]
- The software and chip technology feedback loops in combination could lead to an **AI-technology IE**[13] – so called because AI automates cognitive work in improving software and chip technology, but physical automation isn't required. This would likely be less sudden than a software IE.[14]
- All three feedback loops in combination could lead to a **full-stack IE,**[15] which would likely be the least sudden.

---

11 Even if earlier feedback loops plateau before later feedback loops kick in, it's likely that the later feedback loops will unlock additional gains. For example, better chip designs would create room for software improvements specific to the new chip designs or the new higher quantity of compute.

12 To happen suddenly, a feedback loop would have to both have short time lags and large improvements per iteration. Having short time lags increases the potential for suddenness.

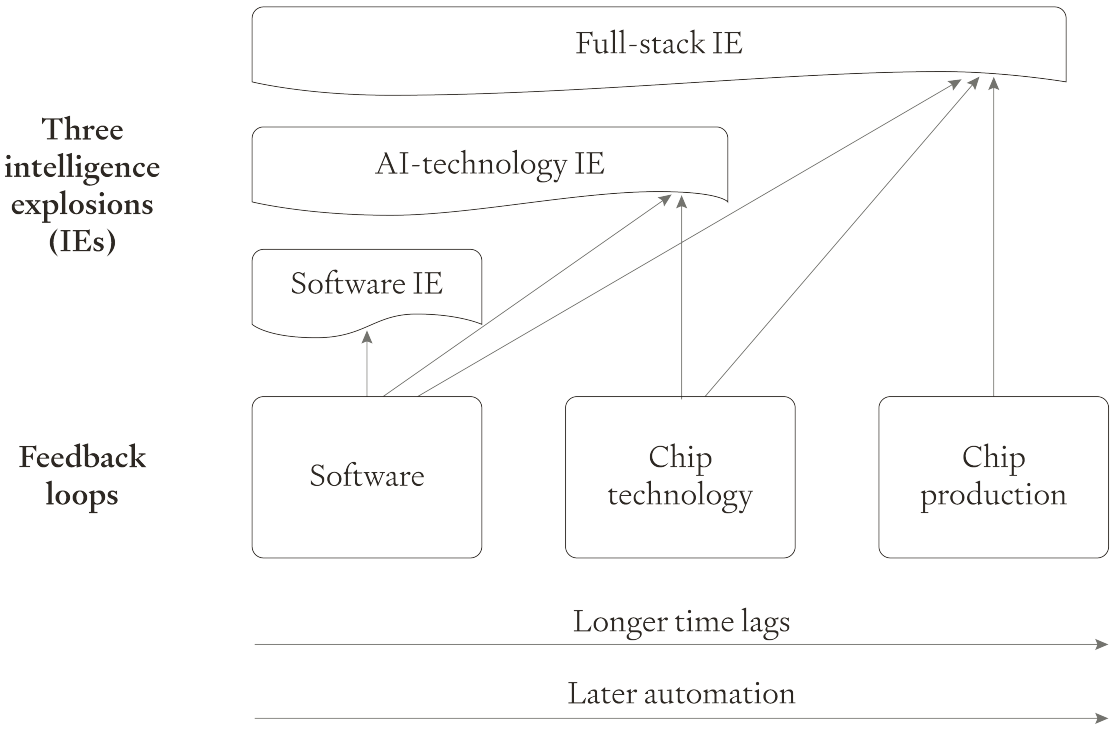
13 If there is a chip technology feedback loop, it's likely that the software feedback loop is also in operation, because:

- Software R&D will likely be automated before hardware R&D.
- Even if the software feedback loop has already slowed by the time the chip technology feedback loop gets going, progress in chip technology will unlock room for further software improvements.

14 This claim could seem confusing at first blush. Adding in the chip technology feedback loop will only make AI progress *faster* and *more sudden* . So why say that the AI-technology IE would likely be *less sudden* than the software IE? We mean that if we condition on there being an AI-technology IE *but there not being a software IE* , we should expect things to be less sudden than if we condition on there being a software IE.

15 If there is a chip production feedback loop in operation, it's likely that both the software and the chip technology feedback loops are in operation too, because:

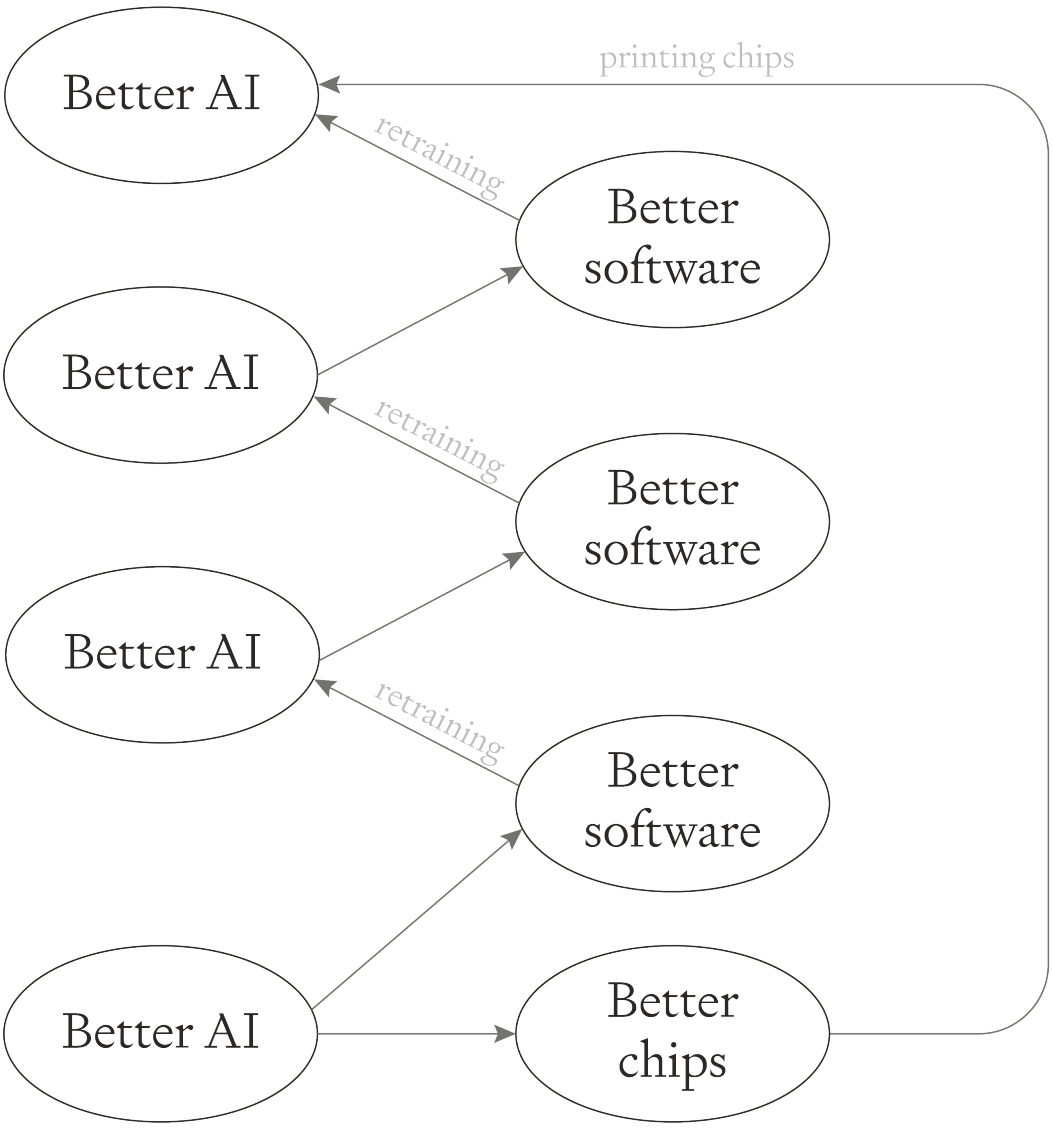
- Software and hardware R&D will likely be automated before hardware production.
- Even if the software and chip technology feedback loops have already slowed by the time the chip production feedback loop gets going, more and better fabs will unlock room for further improvements in chip technology and therefore software.



How the three feedback loops relate to the three intelligence explosions.

These three types of intelligence explosion could all happen in order, or we could see just one or two of them (or none).

This order of intelligence explosions is likely because the relevant feedback loops are likely to be automated in that order. And even if the software and chip technology feedback loops were automated at the same time, the fact that the software loop has shorter time lags means that if there were a software IE, it would still precede an AI-technology IE, as it would take longer for the chip technology loop to feed back on itself.



A software intelligence explosion. Even if AI automates chip R&D at the same time as software R&D, the software feedback loop has a shorter time lag and so could drive the vast majority of AI progress.

We feel the AI-technology and full-stack IEs have been under-analysed.

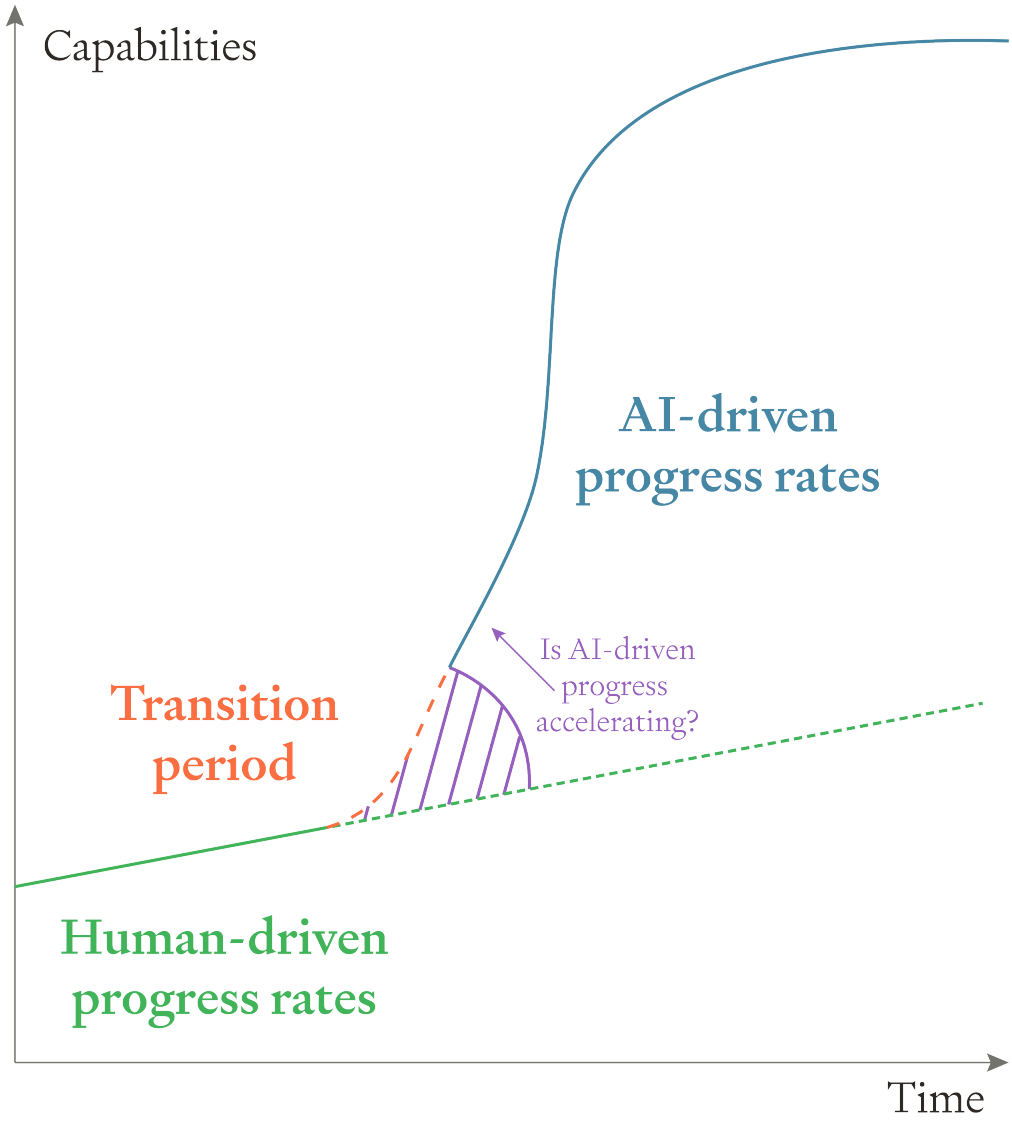
Analysis of these feedback loops

Will AI progress accelerate over time?

Besides time lags, another factor constraining intelligence explosions is that these feedback loops

might be too weak for AI progress to accelerate over time.[16]

When a feedback loop is first automated, there will be an initial speed-up in AI progress. After this point, AI progress may slow back down again if the feedback loop is too weak.



Initially human work drives AI progress (green). Then work to improve AI is gradually automated (orange). Finally, AI systems do almost all the work to improve AI (blue). Empirical evidence suggests that AI progress will (initially) accelerate during this final period.

Whether a feedback loop leads to accelerating progress depends on how much doubling *inputs* to the feedback loop doubles *output* . If output more than doubles when input doubles, there is

16 We define "accelerating AI progress" as "each increment of capability advancement (e.g. GPT-3 → GPT-4) happens more quickly than the last". Historically, each such advancement has required an *exponential* increase in inputs (of compute, data, and human effort), in which case accelerating AI progress would require *super-exponential* growth in inputs.

As we're using the term "intelligence explosion", you could still get an IE without acceleration, if replacing humans with AI made AI progress much faster but progress didn't accelerate further from that point. See here for discussion of initial speedups from AI automation.

accelerating progress (because the *next* doubling of inputs can draw on *more* than twice as many outputs, and so will happen faster than the last doubling).[17]

We can get empirical evidence on this question by looking at the efforts needed to improve AI software and hardware. The key takeaways from our analysis are that, absent human bottlenecks like regulation:

- **A software IE might well accelerate over time,** because the software feedback loop *by itself* might well be enough to sustain accelerating progress (~50% likely).
    - Efficiency gains in various domains of AI suggest that doubling research inputs leads to more than a doubling of compute efficiency (Epoch estimates 0.8 to 3.5 doublings of output per doubling of input across several domains).
    - After considering other kinds of AI progress besides efficiency, and various other adjustments, we think acceleration is fairly plausible (Davidson and Houlden(2025) estimate 1.2 doublings of output for every doubling of cognitive input, with a range of 0.4 to 3.6).
- **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%).
    - Historical data suggests that doubling all hardware R&D inputs has led to ~5 doublings of FLOP/$. Restricting just to cognitive inputs will reduce this number, as will getting closer to effective physical limits - but more than one doubling of output for every doubling of input still seems likely.
- **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 a full-stack IE would accelerate (~90%).
    - If you can build the robots and infrastructure required for building any kind of physical capital, 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 (ignoring resource constraints). If robots also improve robot technology at all, which seems likely, then output would more than double for each doubling of input. 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.

Note that the all-things-considered likelihood of acceleration will be lower, as our analysis sets aside possible human bottlenecks to acceleration, such as regulation or conflict.

This essay presents our analysis in more detail.

17 See here for more explanation on the conditions for accelerating progress.

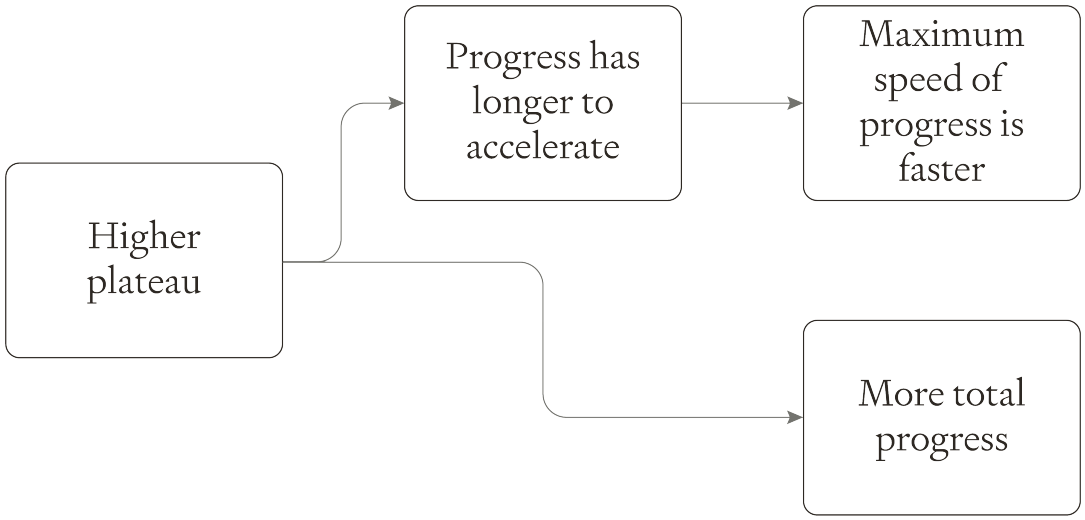
How far could AI progress before hitting effective physical limits?

In addition to whether progress accelerates over time, it is important to consider the *total amount of progress* that each feedback loop can drive before approaching effective physical limits.[18]

If these limits are very high, then not only is there more progress in total, but also *progress has more time to accelerate* and so the maximum speed of progress will be faster.

Theoretical limits for the speed of progress are 100X as fast as recent progress, which would correspond to "effective training compute" doubling every day or so. (For comparison, effective training compute has recently been doubling every ~ 3 months .[19])

The maximum speed could be less than this, for instance due to technical limitations or human bottlenecks. Or it could be slower because physical limits are hit before progress can accelerate to the fastest possible speed. Overall, it's hard to give a good estimate of the fastest speed, but it at least seems plausible that it is very fast.



How much AI progress can we make before hitting effective physical limits? If the limits are higher, more total progress can be made, and the maximum speed of (accelerating) progress will be greater.

How far can each feedback loop progress before hitting physical limits? We can operationalise the distance to physical limits in terms of effective compute.

We estimate that, before hitting physical limits:

- **Software** might improve in efficiency by **~12 OOMs** , with large uncertainty.

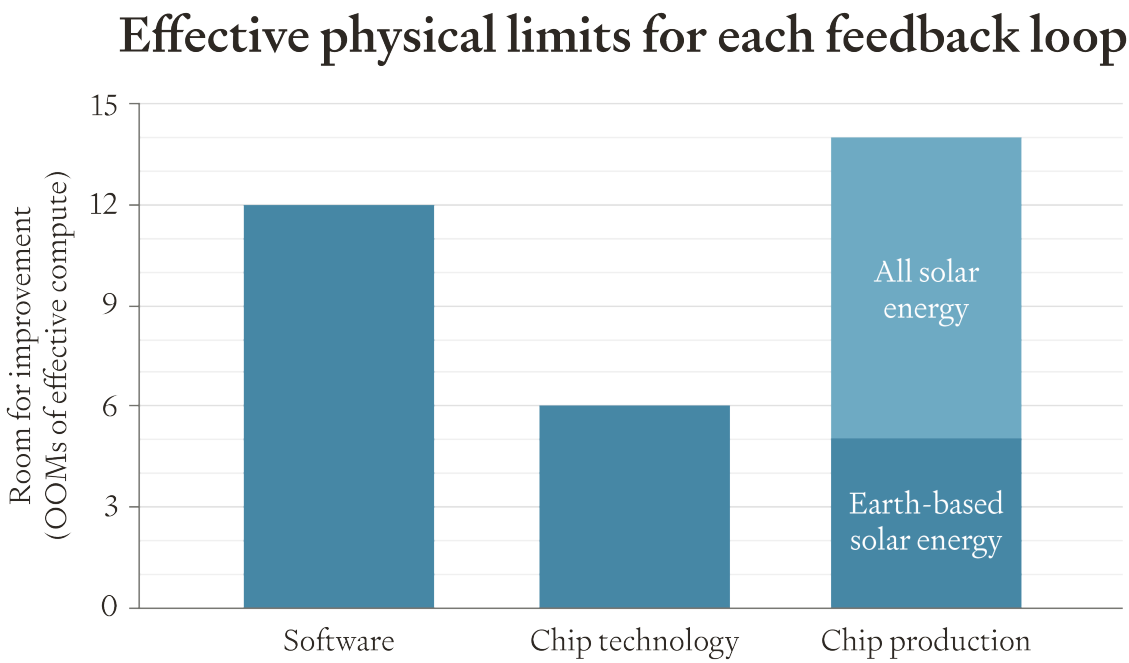
---

18 Of course, progress may slow down long before effective physical limits, because of regulation, diminishing marginal returns or other factors. Here we set aside human constraints and focus on effective physical limits, as these are easier to analyse.

19 See workings here .

- If top-human-level AI is initially trained with 1e29 FLOP, that would be ~5 OOMs less efficient than human learning (which takes ~1e24 FLOP). Then we estimate a further ~7 OOMs of software progress might be possible above the human brain, with very wide uncertainty. (This only includes training efficiency, omitting other sources of software progress.)

- **Chip technology** could likely improve by **~2 OOMs** within the current paradigm, and by a total of **~6 OOMs** if technology approaches Landauer's limit (a physical constraint on the energy efficiency of irreversible computation). Reversible computing could conceivably go further still.
- **Chip production** could scale by **~5 OOMs** using earth-based energy capture, and by a further **~9 OOMs** if space-based solar could capture *all* the energy emitted by the sun.

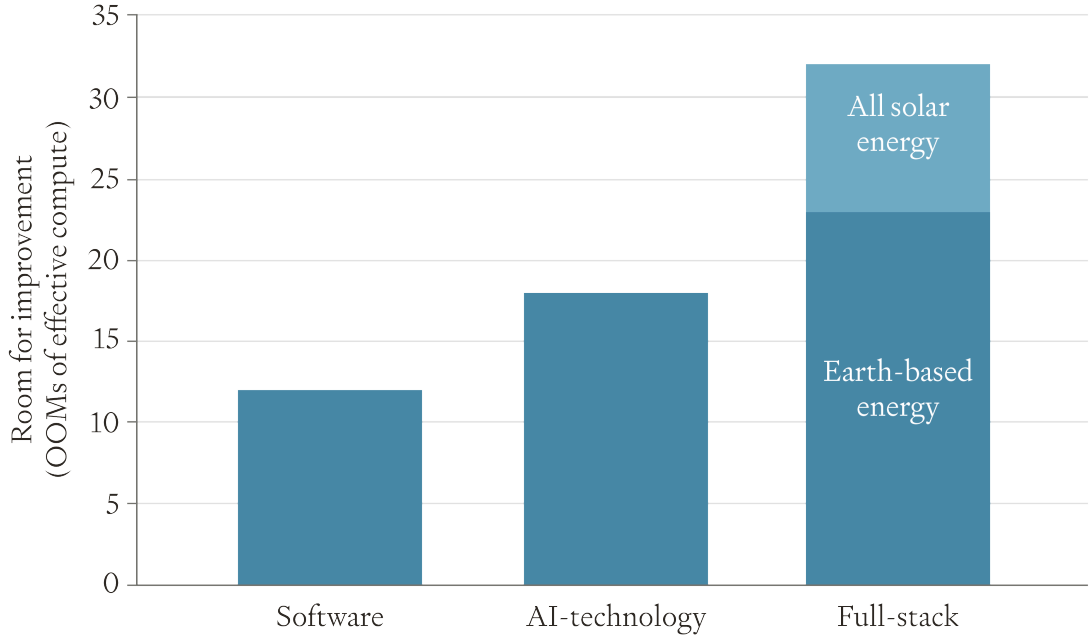


Our estimates of the total room for improvement for each feedback loop before hitting effective physical limits. The limits for software and chip technology might be higher.

To see how far the three intelligence explosions could go before hitting effective physical limits, we can simply add up the limits of each of the feedback loops:

- **The software IE** could increase effective compute by ~12 OOMs, possibly more.
- **The AI-technology IE** could increase effective compute by ~18 OOMs or more.
- **The full-stack IE** could increase effective compute by ~23 OOMs using earth-based energy, or ~32 OOMs using all solar energy.

Effective physical limits for each intelligence explosion



Our estimates of the total room for improvement for each intelligence explosion before hitting effective physical limits. All the limits might be higher.

The landscape of the intelligence explosion

Let's recap our characterisation of the three feedback loops:

Feedback loop	Jobs automated by AI	Time lags in the feedback loop[20]	Likelihood of accelerating progress (ignoring other feedback loops) *(Numbers are very rough!)*	Total room to increase effective compute *(Numbers are very rough!)*
Software	All AI researchers	Training new AI (months),	~50%	~12 OOMs

20 Note that these might shorten during an intelligence explosion, as improved capabilities allow tasks to be completed more quickly (either through having better AI workers, or having more AI workers, or both).

	(e.g. all technical staff at OAI)	though post-training enhancements are faster		
Chip technology	All computing hardware researchers (e.g. all researchers at TSMC, NVIDIA, ASML, etc)	Integrating new tech into chip factories (months) Printing new chips (many months)	~65%	~6 OOMs
Chip production	All labour for building and running chip factories	Building new fabs (years)	~80%	~5 OOMs

This information flows through to characterise the three intelligence explosions:

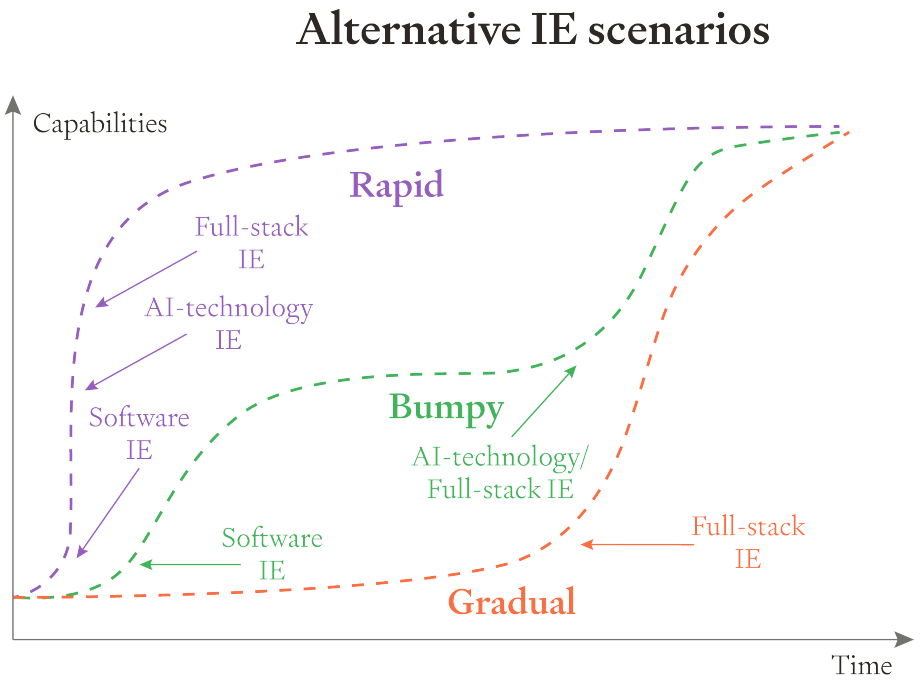
Type of IE	Jobs automated by AI	Time lags in the feedback loop	Likelihood of accelerating progress (ignoring other feedback loops) *(Numbers are very rough!)*	Total room to increase effective compute *(Numbers are very rough!)*
Software	All AI researchers	Training new AI (months), though post-training enhancements are faster	~50%	~12 OOMs
AI-technology	All AI researchers + All computing	Training new AI (months) + Integrating new tech into chip factories	~75%	~18 OOMs

	hardware researchers	(months) Printing new chips (many months)		
Full-stack	All AI researchers + All computing hardware researchers + All physical labour for chip production	Training new AI (months) + Integrating new tech into chip factories (months) Printing new chips (many months) + Building new fabs (years)	~90%	~23 OOMs

Putting all this together, here are three scenarios that we find fairly plausible:

1. **Gradual scenario:** a gradual full-stack IE. The software and chip technology feedback loops alone are not enough to drive accelerating AI progress. There's a full-stack IE that starts slowly (at a similar pace of progress to what we saw in 2020-2024) due to time lags. But it accelerates over time and eventually becomes extremely fast (with a doubling time of months or less) because the effective physical limits are so high.
2. **Bumpy scenario:** a limited software IE followed by slow AI-technology/full-stack IEs. There's a rapid software IE, but it slows down after ~3 OOMs. Later, there are AI-technology and/or full-stack IEs that start somewhat slowly and eventually become extremely fast.
3. **Rapid scenario:**[21] a large software IE (of 6 OOMs or more) followed by fast AI-technology/full-stack IEs. There's a rapid software IE, occurring over months, with fairly high effective physical limits. This improves technology enough to significantly reduce the time lags in the feedback loops for chip technology and chip production. This means the subsequent AI-technology and/or full-stack IEs start out very quickly and before any noticeable slowdown as the software IE starts to plateau.

21 The classic discussion of recursive AI improvements pointed to the possibility of a "fast takeoff", which Nick Bostrom defined as one where the transition "from human-level intelligence to superintelligence" occurs "over some short temporal interval, such as minutes, hours, or days". Bostrom (2014), *Superintelligence: Paths, Dangers, Strategies* , pp. 63-64.
Such a scenario now seems very unlikely, as there will be compute constraints, and as training new models takes ~3 months . Even using very aggressive parameter values in a model of takeoff speeds , takeoff still lasts months. Moreover, arguably the latest language models are already human-level by the standards of this earlier discussion, and we have not yet transitioned to superintelligence.



There has been relatively little strategic thinking about the first two scenarios, in which significantly superhuman AI comes only after long delays and industrial expansion.

Strategic implications for the distribution of power

Different IEs have different implications for the distribution of power, both in terms of which actors will have power over AI development, and how concentrated that power will be.

- A **software IE** would be controlled by the owners of (existing) AI chips and algorithms. Whoever owns the largest stock of AI chips and initially has the best AI algorithms can conduct the faster IE and reach more advanced capabilities first.
    - A software IE would be most likely to occur in the US. Most AI compute is physically located in the US, and most frontier AI developers are US companies.
    - During a software IE, power could become concentrated in just one country, or even in just one company,[22] if the explosion is sufficiently large for the frontrunner to pull very far ahead.
- An **AI-technology IE** would also be directly controlled by the owners of AI chips and algorithms, but owners of the semiconductor supply chain might have significant indirect power, because they control aspects of the chip technology feedback loop.

22 This poses the further risk that the leaders of the frontrunning project leverage AI to perform a coup, which would concentrate power even more.

- As with the software IE, an AI-technology IE would likely involve the US. But continuing the pace of AI progress would also depend on US allies that play a large role in the semiconductor industry, for example Taiwan, South Korea, Japan, and the Netherlands.
- Power would be more broadly distributed than in a software IE, as the semiconductor supply chain would also be critical to AI progress, and it is spread over many countries and companies.

- A **full-stack IE** would be controlled by the owners of AI chips and algorithms, the semiconductor supply chain, and many other parts of the industrial base (including mining, construction, and energy production).
  - Compared to the other two IEs, it is more likely that a full-stack IE would heavily involve countries like China and Saudi Arabia that are willing and able to quickly expand their industry, through a combination of a strong industrial base and a permissive regulatory environment. Authoritarian countries might generally be favoured in this respect over democratic countries.[23]
  - Power would be distributed most broadly during a full-stack IE, as many countries and companies are involved in the chip technology and chip production feedback loops.

Type of IE	Who controls the IE	Which countries are likely involved	How concentrated power will be
Software	Owners of (existing) AI chips and algorithms	The US	Potentially very concentrated
AI-technology	Owners of (existing) AI chips and algorithms, with indirect influence from owners of the semiconductor supply chain	The US and allies (Taiwan, South Korea, Japan, the Netherlands)	More distributed
Full-stack	Owners of AI chips and algorithms, the semiconductor supply chain, and many parts of the industrial base	The US and allies, China, the Gulf states, and many other countries.	Broadly distributed

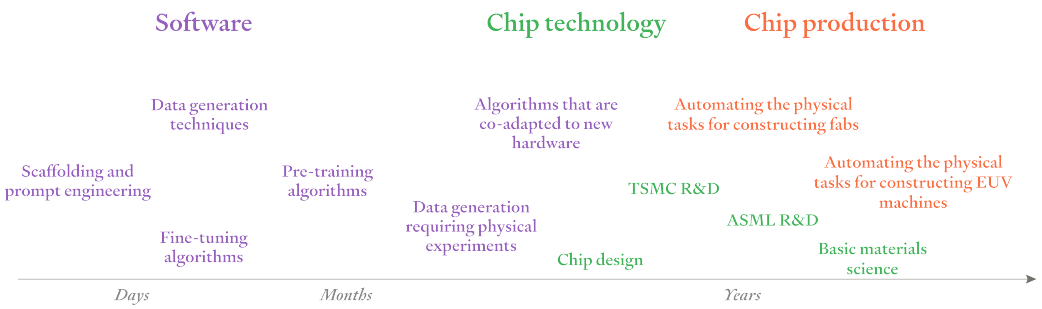
23 Authoritarian countries could also potentially go faster through a full-stack IE as they might be able to maintain higher savings rates than democratic countries.

Thanks to Will Aldred, Adam Bales, Fazl Barez, Stephen Clare, Max Dalton, Oscar Delaney, Daniel Eth, John Haltstead, Fin Moorhouse, Zershaaneh Qureshi, Ben Todd, Lizka Vaintrob and the governing explosive growth seminar group for helpful comments.

Appendix: Are there really just three feedback loops?

The choice of three feedback loops, rather than fewer or more, is natural but somewhat arbitrary.

It's possible to subdivide each of our three feedback loops further.



Time lag (lag from "work done in this area" to "AI is better")

- Within the **software feedback loop**:
  - Post-training (like scaffolding and prompt engineering) could be separated from pre-training.
  - Data generation and fine-tuning learning algorithms could be treated as separate feedback loops (although their time lags are similar).
  - Algorithms that are co-adapted to new hardware and data generation requiring physical experiments would have the longest time lags, and could be separated out.
- Within the **chip technology feedback loop**:
  - Chip design (e.g. by NVIDIA) has much faster turnaround time than ASML's research:
    - It's often possible to start printing the new chip designs without delay.
    - ASML improves EUV machines that must be constructed and integrated into fabs before printing begins.
  - Even more extremely, basic materials science feeds into semiconductor R&D, sometimes with very long time lags. This could be separated out.
- Within the **chip production feedback loop**:

- Automating the final stage of fab construction – putting the final parts together into a fab – has a smaller time lag than automating the whole process of making the machines that make the machines that make the EUV machines.

Some of these time lags are slower than the examples we use in the main text for each feedback loop (and some are faster, like post-training enhancements). The reasons we don't think that the three feedback loops will need to move at the pace of their slowest components are:

- The components with the shortest time lags will get automated first, and generate more and more AI progress. As AI progresses, innovation will reduce the longer time lags.
- In many cases, we think that the slower components are only a small part of what's driving progress:
  - For software, training new models and post-training enhancements seem like larger drivers of progress than physical data generation and co-adapted algorithms.
  - For chip technology, chip design and ASML's research seem like larger drivers of progress than basic materials science.

Appendix: How fast could AI progress eventually become?

The maximum speed of AI progress could conceivably be extremely fast.

We can start by thinking about AI progress in general terms. Epoch estimates that effective compute has been increasing by 10X/year in recent years.[24] This corresponds to 1000X every 3 years (or 150 weeks). But if effective compute were to double in a day, it would only take 10 days (or 1.5 weeks) to 1000X. That's 100X faster than progress today.

Is it possible for effective compute to double that quickly? Here we are reasoning about theoretical limits, rather than giving an all-things-considered estimate. At least for the software and chip production feedback loops, a maximum doubling time of days doesn't seem out of the question:

- **Software** : some post-training enhancements could be implemented and tested in a day. Epoch estimates that selected post-training enhancements are equivalent to increasing effective compute by 5-30X .
  - Doubling times could be faster than a day if AI is able to find better post-training enhancements.
  - At some point we will reach diminishing marginal returns on post-training enhancements. Then the limit on the rate of software progress would become the time taken to train new models. This is currently months, though it could get shorter with breakthroughs in software or chips.
- **Chip technology** : it's hard to know if chip technology could double this fast.

24 4.2X compute and 3X algorithmic improvements.

    - There are currently longer time lags for chip technology than for software, because of the time it takes to print chips. But if chips can be made in a day (see below), there's no obvious reason that chip technology couldn't double as fast as software and chip production.
- **Chip production** : fruit flies can double in days.[25] This is proof of concept that biological replicators can double their compute (i.e. brains) in days.
    - It doesn't clearly follow that artificial replicators can make general-purpose computer chips as fast as fruit flies can make brains. In particular, the computing power in a fruit fly brain cannot flexibly run many different types of software like computer chips can. It might take more time to make flexible computing power.[26]
    - On the other hand, it would be surprising if evolution had found the optimal configuration for fast replication of compute, and machines designed to replicate as fast as possible may well exceed biological limits.

There are several reasons that maximum speed could be less than this. Human bottlenecks like regulation could slow doubling times down. Or we could approach physical limits before progress can accelerate to the maximum possible speed.

Overall, it's hard to give a good estimate of the fastest speed, but it at least seems plausible that it is very fast. **It's conceivable that maximum AI progress could become 100X as fast as recent progress (or faster), with effective compute doubling every day or so.**

25 At 29°C, it takes 7 days from egg to adult, and total lifespan is around 40 days . Females can lay up to 100 eggs per day, from the day they become adults. Naively, a population of 2 fruit flies (one male and one female) would become a population of 202 adults in 8 days (~1 day doubling time).

26 Computer chips can execute any software that is written in machine-readable code. This allows anyone to write software to their liking, translate it to machine-readable code, and then use the computer chips to run the software. This isn't possible with fruit flies because there is no "fruit-fly-readable code" – no flexible way to describe software that the fruit flies' brain can interpret.