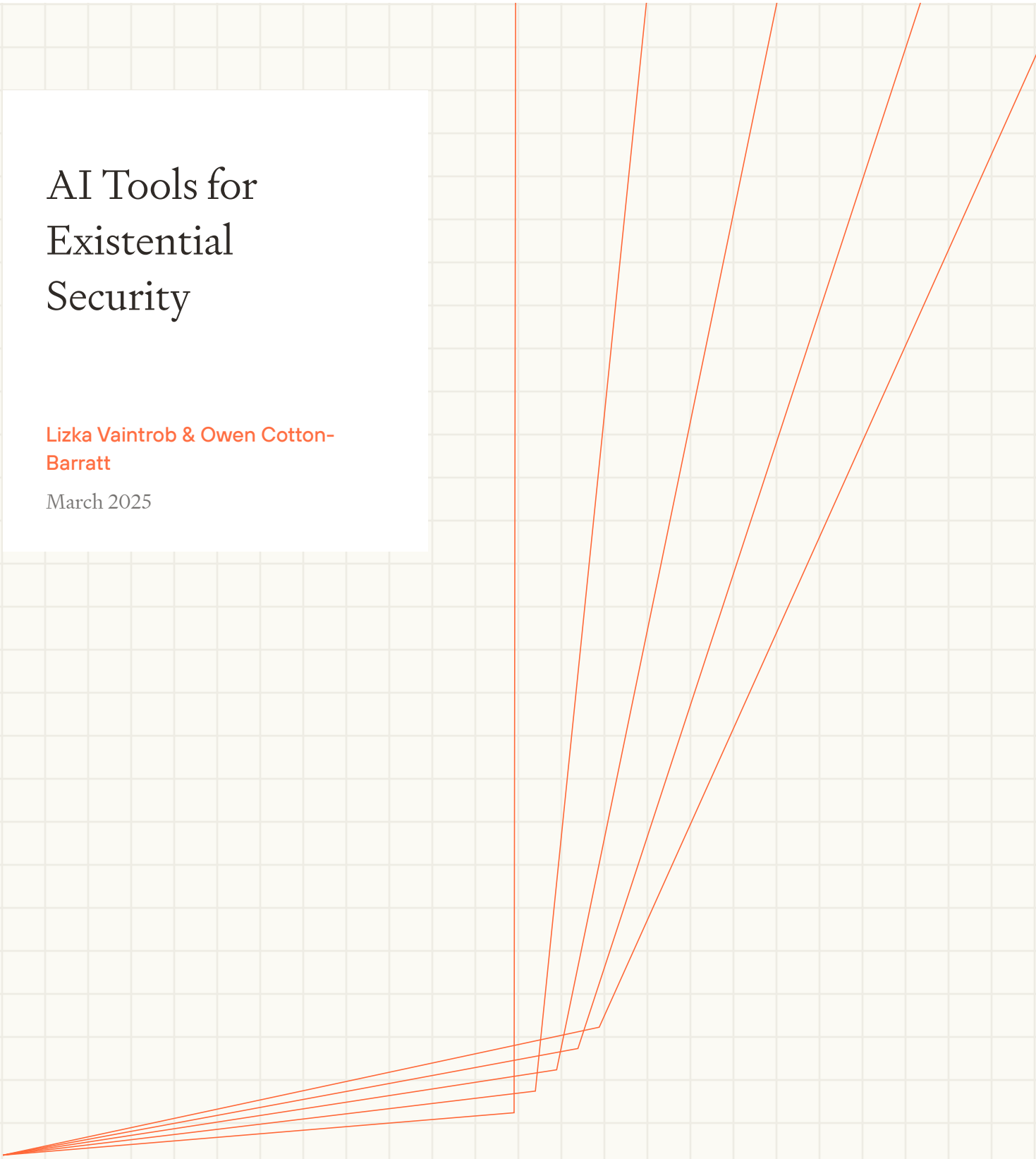
Forethought

AI Tools for Existential Security

Lizka Vaintrob & Owen Cotton-Barratt

March 2025



Contents

AI Tools for Existential Security

Lizka Vaintrob & Owen Cotton-Barratt

We're grateful to Max Dalton, Will MacAskill, Raymond Douglas, Lukas Finnveden, Tom Davidson, Joe Carlsmith, Vishal Maini, Adam Bales, Andreas Stuhlmüller, Fin Moorhouse, Davidad, Rose Hadshar, Nate Thomas, Toby Ord, Ryan Greenblatt, Eric Drexler, and many others for comments on earlier drafts and conversations that led to this work. Owen's work was supported by the Future of Life Foundation.

For correspondence, please email contact@forethought.org.

Rapid AI progress is the greatest driver of existential risk in the world today. But — if handled correctly — it could also empower humanity to face these challenges.

Executive summary

1. Some AI applications will be powerful tools for navigating existential risks
Three clusters of applications are especially promising:

- ***Epistemic*** **applications** to help us anticipate and plan for emerging challenges
  - e.g. high-quality AI assistants could prevent catastrophic decisions by helping us make sense of rapidly evolving situations
- ***Coordination-enabling*** **applications** to help diverse groups work together towards shared goals
  - e.g. automated negotiation could help labs and nations to find and commit to mutually desirable alternatives to racing
- ***Risk-targeted*** **applications** to address specific challenges
  - e.g. automating alignment research could make the difference between "It's functionally impossible to bring alignment up to the requisite standard in time" and "this is just an issue of devoting enough compute to it"

2. We can accelerate these tools instead of waiting for them to emerge

- While broad AI progress will drive the development of many applications, we have some flexibility in the timing of specific applications — and even small speed-ups could be crucial (e.g. by switching the order of risk-generating capabilities and risk-reducing ones)
- We could use a variety of strategies to accelerate beneficial applications:
  - ***Data pipelines & scaffolding*** : by curating datasets or scaffolding for key capabilities, or laying the groundwork to automate this, we could enable those capabilities as soon as underlying AI progress supports them
  - ***Complementary tech & removing other barriers to adoption*** : by building out the UI or other complementary technology, and ensuring that people are eager to use the applications, we could enable the applications to see use as soon as the underlying capabilities are there, rather than accept delays to adoption
  - ***Shaping compute allocation*** : by building support among key decision-makers who might allocate compute, we could ensure that crucial applications are among the earliest to see large amounts of automated research
- Accelerating beneficial applications can often be done unilaterally (in contrast to delaying dangerous capabilities, which may need consensus)

Implications

These opportunities seem undervalued in existential risk work. We think a lot more people should work on this — and the broader "differential AI development" space. Our recommendations:

1. **Shift towards accelerating important AI applications**
    - e.g. curate datasets for automating alignment research; or build AI forecasting systems
2. **Plan for a world with abundant cognition**
    - Some new approaches will come online, and some current work may be obsoleted
    - e.g. it could make sense to build tools that process rich information to provide bespoke infectious disease exposure advice in contact tracing apps
3. **Get ready to help with automation**
    - e.g. build relevant expertise, or work towards institutional buy-in

Some AI applications will help navigate existential risks

Epistemic applications

People are more likely to handle novel challenges well if they can see them coming clearly, and have good ideas about what could be done about them.

Examples of promising *epistemic* applications[1]

Applications	How they might help
AI forecasting tools	High quality forecasts, especially of novel technological developments and their strategic implications, could help us to anticipate and prepare for key challenges. Sufficiently trusted AI systems with strong general track records could help to align expectations between parties.
AI for collective epistemics	AI systems that do high-quality fact-checking (or evaluate other systems for how truthful or enlightening they are) could help people to stay oriented to what is reliable in the world, and avoid failures of coordination from misplaced trust.
AI for philosophy	By helping people to engage in moral reflection, or directly tackling hard philosophical questions, AI systems might help humanity to avoid subtle but catastrophic *moral* errors.

1 See e.g. Lukas Finnveden's post on AI for epistemics for further discussion of this area.

	Moreover, poor philosophical grounding could lead superintelligent AI systems to go off the rails in some ways.

Coordination-enabling applications

Local incentives sometimes prevent groups from achieving outcomes that would benefit everyone. This may make navigating key challenges — for example, coordinating to go slow enough with AI development that we can be justifiably confident it is safe — extremely difficult. Some AI applications could help people to coordinate and avoid such failures.

Examples of promising *coordination-enabling* applications:

Applications	How they might help
Automated negotiation tools	Negotiation processes often fail to find the best mutually-desirable outcomes — especially when time is limited, there are many parties involved, or when it's hard to exchange information openly. AI tools could relieve bandwidth issues, or permit the perfectly confidential processing of relevant private information.
Automated treaty verification and/or enforcement tools	In some cases, all decision-makers would be happy with a potential agreement if they could trust that everyone would follow it, but trust issues prevent the agreement. Verification systems can mitigate the issue, and AI systems can improve them (e.g. by improving monitoring systems, or by serving as arms inspectors who could be trusted not to leak sensitive information). Sufficiently robust AI systems could even be empowered to enforce certain treaty provisions.
AI tools for structured transparency	Technological progress may lead to a position where it is easy to construct extremely destructive weapons . AI monitoring could ensure, for instance, that people weren't building weapons or help developers understand how people are using advanced AI models — without

	creating the privacy issues normally associated with surveillance.

Better coordination tools also have the potential to cause harm. Notably, some tools could empower small cliques to gain and maintain power at the expense of the rest of society. And commitment tools in particular are potentially dangerous, if they lead to a race to extort opposition by credibly threatening harm; or if humanity "locks in" certain choices before we are really wise enough to choose correctly.[2]

Risk-targeted applications

Examples of promising *risk-targeted* applications:

Applications	How they might help
Automating research into AI safety, such as theoretical alignment, mechanistic interpretability, or AI control	If these areas are automated early enough relative to the automation of research into AI capabilities, safety techniques might keep up with increasingly complex systems. This could make the difference in whether we lose control of the world to misaligned power-seeking AI systems.[3]
AI tools for greatly improving information security	Strong information security could limit the proliferation of powerful AI models, which could facilitate coordinating not to race forwards as fast as possible. It could also reduce the risk of rogue models self-exfiltrating.
AI-enabled monitoring systems for pandemic pathogens	Screening systems could prevent malicious actors from synthesizing new pandemic-capable viruses. AI-assisted biosurveillance could detect transmission of threatening viruses early enough to contain them.

2 There is a bit more discussion of potential downsides in section 5 of this paper: here .

3 This strategy has been discussed in many places, e.g. here , here , here , and here .

Other applications?

Applications outside of these three categories might still meaningfully help. For instance, if food insecurity increases the risk of war, and war increases the risk of existential catastrophe, then AI applications that boost crop production might indirectly lower existential risk.

But we guess that the highest priority applications will fall into the categories listed above[4] , each of which focuses on a crucial step for navigating looming risks and opportunities:

- Identifying the challenges we're facing and which strategies could help navigate them;
- Coordinating on those strategies;
- And actually implementing the strategies.

We can accelerate helpful AI tools

There's meaningful room to accelerate some applications

To some extent, market forces will ensure that valuable AI applications are developed not too long after they become viable. It's hard to imagine counterfactually moving a key application forward by decades.

But **the market has gaps, and needs time to work.** AI is a growth industry — lots of money and talent is flowing in *because* the available opportunities exceed the degree to which they're already being taken. So we should expect there to be some room to counterfactually accelerate any given application by shifting undersupplied capital and labour towards it.

In some cases this room might be only a few months or weeks. This is especially likely for the most obviously economically valuable applications, or those which are "in vogue". Other applications may be less incentivized, harder to envision, or blocked by other constraints. It may be possible to accelerate these by many months or even years.

Moreover, **minor differences in timing could be significant.** Even if the speed-up we achieve is relatively small or the period during which the effects of our speed-up persist is short, the effects could matter a lot.

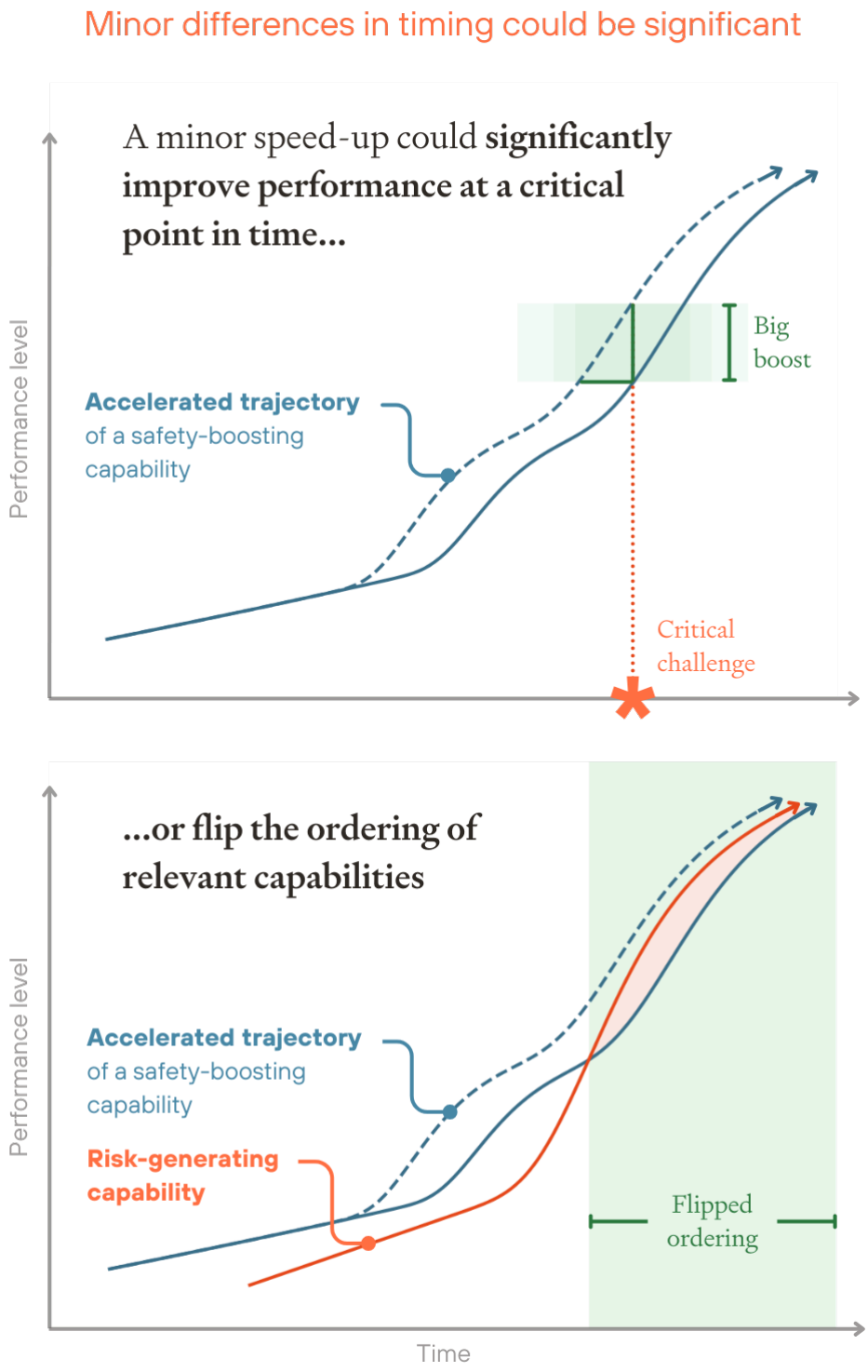
This is because, at a time of rapid progress in AI:

1. Small differences in time could represent major differences in capability level
    - What is "technologically feasible" given general AI capabilities may be shifting rapidly, so that keeping up with that frontier instead of lagging months behind it could mean a big difference in practice
2. Small boosts could flip the ordering of key capabilities

---

4 Not that there is anything definitive about this categorization; we'd encourage people to think about what's crucial from a variety of different angles.

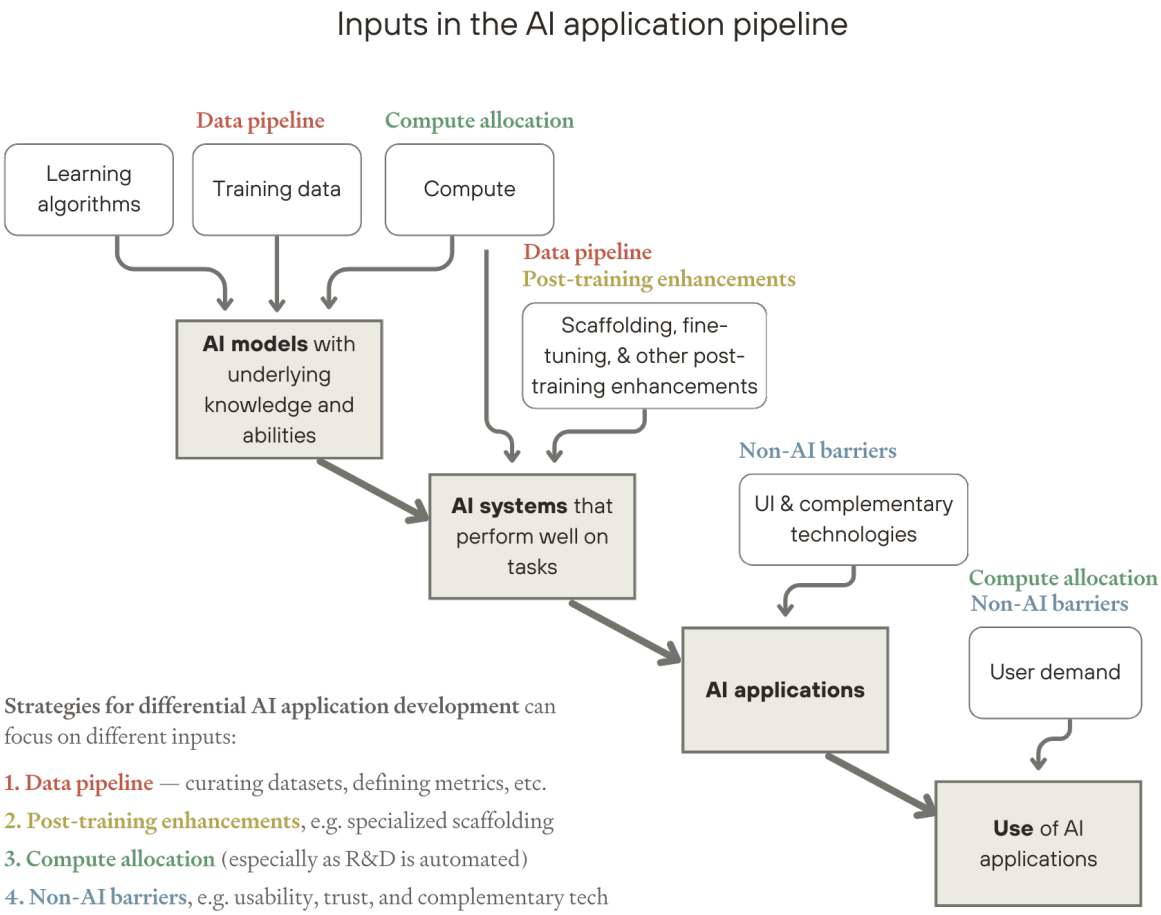
- Achieving risk-reducing capabilities before[5] the risk-generating capabilities they correspond to could have a big impact on outcomes



5 And note it's not just about the ordering of the capabilities, but about whether we have them in a timely fashion so that systems that need to be built on top of them actually get built.

There are promising strategies for accelerating specific AI applications

We have promising strategies that focus on almost all[6] of the major inputs in the development of an AI application.



1. Invest in the data pipeline (including task-evaluation)

High-quality task-specific data is crucial for training AI models and improving their performance on specific tasks (e.g. via fine-tuning), and it's hard to get high-quality data (or other training signals) in some areas. So it could be very useful to:

- Curate specialized datasets
    - e.g. if we want to accelerate AI systems which can support people in avoiding making big mistakes in their decision-making, maybe we should collect datasets of identified errors
    - e.g. if we want to speed up automation of research in a given field, maybe we should try to collect and share more intermediate research products (working notes, conversations) for future training, or build infrastructure to do this
- Define robust task-evaluation schemes

6 The main exceptions are learning algorithms and in most cases architectures, which are typically too general to differentially accelerate specific applications.

- Metrics (benchmarks or other ways to grade performance) tend to accelerate the development of systems that perform well on the associated tasks — principally because they can enable rapid, automated learning (as in the case of self-play for AlphaGo), but also because they can become targets for developers
- So it may be high-leverage to develop evaluation schemes for performance on tasks we care about[7]
- e.g. if we want to improve automated negotiation tools, we might invest in benchmarks for assessing the quality of tool performance

2. Work on scaffolding and post-training enhancements

Techniques like scaffolding can significantly boost pre-trained models' performance on specific tasks. And even if the resulting improvement is destined to be made obsolete by the next generation of models, the investment could be worth it if the boost falls during a critical period or create compounding benefits (e.g. via enabling faster production of high-quality task-relevant data).

3. Shape the allocation of compute

As R&D is automated, choices about where compute is spent will increasingly determine the rate of progress on different applications.[8] Indeed, under inference paradigms this is true more broadly than just for R&D — larger compute investment may give better application performance. This means it could be very valuable to get AI company leadership, governments, or other influential actors on board with investing in key applications.

4. Address non-AI barriers

For some applications, the main bottleneck to adoption won't be related to underlying AI technologies. Instead of focusing on AI systems, it might make sense to:

- Improve user interfaces or build a more accessible version of an existing application
- Increase demand, for instance by working on reputational issues for a given application (e.g. just building a trusted brand!), or ensuring that key institutions are prepared to adopt important applications quickly
- Develop complementary technologies (e.g. some privacy-conscious actors might use certain AI applications only with sufficiently good privacy systems)

Different situations will call for different strategies. The best approach will be determined by:

- The likeliest bottlenecks for a given application
- The target timeline (some strategies take longer to pay off or help at different technological levels)
- The levers that are available

---

7 In some cases, optimizing for a task metric may result in spillover capabilities on other tasks. The ideal metric from a differential acceleration perspective is one which has less of this property; although some spillover doesn't preclude getting differential benefits at the targeted task.

8 AI companies are already spending compute on things like generating datasets to train or fine-tune models with desired properties and RL for improving performance in specific areas. As more of AI R&D is automated (and changing research priorities becomes as easy as shifting compute spending), key decision-makers will have more influence and fine-grained control on the direction of AI progress.

The most effective implementation of one of these strategies won't always be the most direct one. For instance, if high-quality data is the key bottleneck, setting up a prize for better benchmarks might be more valuable than directly collecting the data. But sometimes the best approach for accelerating an application further down the line will involve simply building or improving near-term versions of the application, to encourage more investment.

These methods can generally be pursued unilaterally. In contrast, *delaying* an application that you think is harmful might more frequently require building consensus. (We discuss this in more detail in an appendix.)

Implications for work on existential risk reduction

Five years ago, working on accelerating AI applications in a targeted way would have seemed like a stretch. Today, it seems like a realistic and viable option. For the systems of tomorrow, we suspect it will seem obvious — and we'll wish that we'd started sooner.

The existential risk community has started recognizing this shift, but we don't think it's been properly priced in.

This is an important opportunity — as argued above, some AI applications will help navigate existential risks and can be meaningfully accelerated — and it seems more tractable than much other work. Moreover, as AI capabilities rise, AI systems will be responsible for increasing fractions of important work — likely at some point a clear majority. Shaping those systems to be doing more useful work seems like a valuable (and increasingly valuable) opportunity for which we should begin preparing for now.

We think many people focused on existential risk reduction should move into this area. Compared to direct technical interventions, we think this will often be higher leverage because of the opportunity to help direct much larger quantities of cognitive labour, and because it is under-explored relative to its importance. Compared to more political interventions, it seems easier for many people to contribute productively in this area, since they can work in parallel rather than jostling for position around a small number of important levers.[9] By the time these applications are a big deal, we think it could easily make sense for more than half of the people focusing on existential risk working on related projects. And given how quickly capabilities seem to be advancing, and the benefits of being in a field early, we think a significant fraction — perhaps around 30% — of people in the existential risk field should be making this a focus today.

What might this mean, in practice?

9 This work may also be more promising than policy-oriented work if progress in AI capabilities outpaces governments' ability to respond

1. Shift towards accelerating important AI applications

Speed up AI for current existential security projects

If you're tackling an important problem[10] , consider how future AI applications could transform your work. There might already be some benefits to using AI[11] — and using AI applications earlier than might seem immediately useful could help you to learn how to automate the work more quickly. You could also take direct steps to speed up automation in your area, by:

- Identifying good tasks to automate — ones that are:
  - Worth scaling up (not rare, one-off tasks)
  - Self-contained, with clear inputs and outputs
  - Within reach
- Redesigning your processes to be automation-friendly — using standard templates, clear documentation, etc.
- Gathering and documenting examples of good work for training current or future AI systems

Work on new AI projects for existential security

You might also accelerate important AI applications by:

- Starting or joining new projects (e.g. startups) that are building those applications
- Joining or partnering existing institutions which are holding key functions (e.g. democratic oversight), and helping them with automation

2. Plan for a world with abundant cognition

As AI automates more cognitive tasks, strategies that were once impractically labour-intensive may become viable. We should look for approaches that scale with more cognitive power, or use its abundance to bypass other bottlenecks.

Newly-viable strategies might include, e.g.:

- In epistemic tools — automatically propagating updates from each new piece of information through a knowledge database, to highlight places where it might require a significant rethink of strategy
- In coordination tools — exploring a large range of possible agreements, and examining what different potential coalitions might think of them, in order to identify the best directions to go in

---

10 Although you should also be conscious that "which problems are important" may be changing fairly rapidly!

11 As we write this (in February 2025) we suspect that a lot of work is on the cusp of automation — not that there are obvious huge returns to automation, but that there are some, and they're getting bigger over time.

- In biosecurity — processing rich information on a bespoke person-by-person basis, to enable significantly better contact tracing apps
- In AI safety — modeling and assessing the alignment impacts of different possible updates to model weights, and using these assessments to make smarter updates than blind gradient descent

The other side of this coin is that some current work is likely to soon be obsolete. When it's a realistic option to just wait and have it done cheaply later, that could let us focus on other things in the short term.

3. Get ready to help with automation

Our readiness-to-automate isn't a fixed variable. If automation is important — and getting more so — then helping to ensure that the ecosystem as a whole is prepared for it is a high priority.

This could include:

- Positioning yourself in the job market, by joining relevant startups or existing institutions that will need to automate things well
- Getting experience working with state-of-the-art AI tools, and using them to do things
- Developing deep expertise about domains that may be particularly valuable to automate
- Investing further in strategy and prioritization to help the field as a whole to orient
- Creating infrastructure to help people coordinate and share knowledge about how best to automate high-value areas
- Fostering understanding, buy-in, or community among relevant people

Further context

In appendices, we discuss:

- Whether accelerating applications could be bad via speeding up AI progress in general
- Dynamics that might make it harder to meaningfully counterfactually accelerate an application (and when meaningful acceleration still looks achievable)
- How our proposal relates to existing concepts, such as def/acc, differential technological development, and differential *AI* development
- Within "differential AI development," the distinction between differential *application* development and differential *paradigm* development, and the relative advantages of each
- Why we've focused on accelerating risk-reducing applications rather than slowing down risky applications