

Merge with this live article into here when this is finished. Possibly rename this one

☰ Why should we prepare for human-level AI technology now rather than decades down...

It has been argued that it is currently too early to work on AI safety, both because the threat is not imminent and because we don't yet have access to models which we know are dangerous if misaligned. For instance, Yann LeCun [claims](#) that worrying about AI safety now is like worrying about turbojet engine safety in 1920.

There is [substantial uncertainty](#) as to when AGI is coming. If it is coming very soon, the time to work on safety is definitely now¹. But what if, for the sake of argument, you think AGI is still decades away? Even then, there are still many reasons to work on safety now:

- We can make progress on alignment before we have AGI
- It's hard to know when AGI will be imminent and the present time might not be particularly early
- Adding safety as an afterthought does not yield good results
- Safety solutions should be tested by time
- Social and political preparations take time
- An unknown amount of fundamental work might be needed

First, it's worth noting that it is not necessary to wait for AGI to make progress on aligning it.

- Progress on [agent foundations](#) does not depend on the capabilities of current AI.
- Most interpretability work is done on toy models, which are much less capable than current day foundation models.²
- There are [known examples](#) of specification gaming with current systems which computer scientists do not currently know how to solve in a systematic way.
- The invention of RLHF [in 2015](#), years before it was first used in LLMs, illustrates that conceptual breakthroughs can happen before the technology to use them is available.
- We can develop [general methods for identifying undesirable behavior](#) before we build the systems we want to test.

It is also difficult to know when AGI is near, and there's [no universal agreement on what would signal imminent AGI](#). We also don't know how much more work needs to be done to align AI, and some argue that we still need to make fundamental breakthroughs before it is possible.

One reason to avoid waiting is that adding safety to a system as an afterthought does not yield good results. [Hendrycks et al.](#) explain that "if attention to safety is delayed, its impact is limited, as unsafe design choices become deeply embedded into the system," citing a report for the Department of Defense which concludes that "approximately 75% of the most critical decisions that determine a system's safety occur early in development". They

¹ If we live in such a world, the time to start seriously working on this was probably 10 years ago.

² OpenAI attempted to use [GPT-4 to interpret the much simpler GPT-2](#) and Anthropic's work on [monosemanticity](#) is done on a toy model with only one layer.

mention the internet as an example of a system which remains unsafe decades after it was built because it was not built to be safe.





Another reason is that it takes time to make sure that safety solutions work. Hendrycks et al. argues that expert validation is insufficient, as well-regarded solutions can have hidden flaws. They cite the example of the [four color theorem](#), where a flaw in a peer-reviewed proof remained undetected for years, and a correct proof took almost a century more. Doing machine learning safety research early can provide more time for people to check solutions, and reduce the likelihood of accidents.

In addition to technical preparations, social and political preparations also take time. In making decisions about AI, we face the [Collingridge dilemma](#): if we wait to see how it impacts society, it may become deeply embedded and hard to change. It takes time to properly formulate and build consensus around regulation, and this process has to be completed before harms become entrenched.

Alternative phrasings

- Why research alignment in advance of developing the systems for which we need it?
- Is it possible to work on aligning AGI without first building AGI?

Related

-  Isn't AGI decades away?
-  How long will it be until human-level AI is created?
-  What is the sharp left turn?
-  What is a "treacherous turn"?

Scratchpad

Current structure (according to Murphart):

- Safety cannot be an afterthought (Hendrycks citation)
- We want solution before AGI, but AGI date unknown
- There might be an unknown amount of fundamental work needed before we can align AGI
- This might not be particularly early
- Some relevant work on alignment has been done years before it was used (Yaakov is uncertain if we should keep this)
- Social and political preparations also take time

From Yoshua Bengio's [FAQ](#)

“Q. Since we do not understand yet exactly what a superhuman AI would look like, it is a waste of time to try to prevent such unknown risks. Could we have figured airplane safety rules before the Wright brothers? Let’s fix the problems with very powerful AI systems when we will understand them better.”

“A: I used to think exactly like this, thinking that superhuman intelligence was still far in the future, but ChatGPT and GPT-4 have considerably reduced my prediction horizon (from 20 to 100 years to 5 to 20 years). With over 100 million users, we are well past the Wright brothers stage. These LLMs have also given us pretty good clues of what an AI could already do and what it is missing, and several research groups are working on these shortcomings. The unexpected speed at which LLMs have acquired their current level of competence simply because of scale suggests that we could also see the rest of the gap being filled in just a few years with minor algorithmic changes. Even if someone disagrees with the temporal horizon distribution, I don’t see how one could reject that possibility. I acknowledge your argument that it is difficult to come up with regulation and countermeasures for something that does not yet exist. However, there are examples of proposals to control dangerous technologies (including atomic power in the 1910s and AI in this century, or biological agents regulated with a global regime that is agnostic to the exact pathogens that could be used) that have been made that did not rely on knowing the exact form of the technology. The other important element here is the slowness of adaptation of society, not to mention of governments to implement policies and regulations. I believe that we should study and evaluate preventative measures that we could take as a society to reduce those risks and gradually prepare countermeasures, and we should get started as soon as possible. Generic policies, like [monitoring and capability evaluation](#), licensing, reporting requirements and [auditing](#) of dangerous technologies, are applicable for all technologies. See also [this discussion](#) on the diversity of actions one should consider to mitigate catastrophic AI risks. Our lack of understanding and visibility of harm scenarios indeed poses difficult dilemmas regarding regulation, though (e.g. see the [Collingridge dilemma](#)). Finally, going back to how a superhuman AI might look like, there is already a working hypothesis: take the current generative AI architectures and train them (as inference machines, see [this blog post](#)) with system 2 machinery and objectives (which admittedly need to be scaled up) so that they can also reason better, be more coherent and imagine plans and counterfactuals. It would still be a big neural net trained with some objective function and some procedure for generating examples (not just the observed data). We now have a lot of experience with such systems, and there are many open research questions about how to make them safe and trustworthy.”

“A: In addition, even if we do not fully master all the principles that explain our own intelligence (i.e. systems 1 and 2), digital computing technology can bring additional advantages over biological intelligence. For example, computers can parallelize learning across many machines thanks to high-bandwidth communication enabling them to exchange trillions of model parameters, while humans are limited to exchanging information at the rate of a few bits per second via language. As a result, computers can learn from much larger datasets (e.g. reading the whole internet) which is infeasible for humans in-lifetime – see [Geoff Hinton’s arguments](#) along this line, especially starting around 21m37s.”

“A: Finally, even if an AI is not stronger than humans on all cognitive abilities, it could still be dangerous if the aspects it masters (e.g. language but not robotics) are sufficient to wreak havoc, for example, using dialogue with humans to create a manipulative emotional connection and pay or influence them to act in the world in ways that could be very harmful, starting with destabilizing democracy even more than current social media. We know that at least a subset of humans are very suggestible and can believe, for example, conspiracy theories with a conviction that is greatly out of proportion to their evidence. In addition, organized crime is likely to execute well-paid tasks without even knowing that they are being paid by an AI.”

end of quote from yoshua bengio

Old drafts (by Rory)

There may be two options: work on AI safety early, or end up dead.

This claim might sound hyperbolic, but it is seriously entertained by a team of researchers at DeepMind, who worry that AI development may lead us to encounter a [sharp left turn](#).

Here’s the thought: we initially develop AI that accomplishes “impressive feats”, whose skills appear to generalize well to novel domains which weren’t part of its original training. We can already see this to a limited degree with GPT-3, which is able to write poetry and produce code — despite not initially being trained to do so. The thought behind the idea of a sharp left turn is that, while AI *capabilities* will generalize to novel domains, *alignment techniques* will fail to generalize (at least to the same degree). Moreover, once AIs move beyond a certain level of capability, humans will no longer be able to intervene to ensure alignment.

While we don’t know for sure whether we’ll encounter a sharp left turn, various pieces of evidence suggest that we might be in a situation where capabilities are more likely to generalize than alignment.

1. First, there are likely to be greater incentives towards capabilities than alignment. Firms may face [competitive pressures to sacrifice safety for speed](#) of capability improvements, as research organizations view themselves as engaged in a [collective action problem](#). This suggests that society will *underinvest* in safety techniques relative to capabilities.
2. Second, AIs – once past a certain point of capability – are likely to be averse to mechanisms which tamper with their values. There are incentives which favor the development of more powerful capabilities, and incentives which – beyond a certain point – actively *disfavor* value change.
 - a. This suggests that we should work on alignment before we encounter AIs approaching anywhere close to human capabilities. If we don’t put active effort in to produce *robustly generalizable* alignment strategies, a reasonable default is to expect capability generalization, but not *alignment* generalization.

If alignment research were either easy or (comparatively) well-resourced, we might not need to worry about working on AI safety early, as we could be confident that robust alignment techniques would emerge when we needed them. However, alignment research is also *neglected*, and (depending on your timelines) potentially *urgent*. While global investment into AI hovers around [\\$100 billion per year](#), spending on AI safety is less than \$100 million — or 0.1% of total AI spending.

Moreover, ensuring alignment might be extremely *difficult*. If a powerful optimization process selects for some goal, there's no guarantee that the goal will generalize. Humans in the evolutionary environment were selected for inclusive genetic fitness, whereas humans today have more complicated values — values very far removed from optimizing for inclusive genetic fitness. Even in comparatively limited AI systems today, we see (weak) empirical support for [capabilities generalizing further than alignment](#).

While a 'sharp left turn' makes earlier work on AI safety more urgent, we should note that believing in this view is not the *only* reason one might have for promoting early work on AI safety. For instance, you might believe that advanced AI – whenever it arrives, and whatever its dynamics – is likely to be a pivotal technology for the future. In this case, working on AI safety early may help influence the degree to which [AI safety maintains itself as a more mainstream and acceptable issue in the future](#).

Some points from this paper on [unsolved problems in AI safety](#):

A study in the military was conducted which showed that 75% of the critical decisions which would later define the safety and security of a system were made early on in the development of that system. An example of this is the internet which was made mostly for use by academics. Little thought was given to how to make the early internet secure, and now decades later making the internet secure is extremely difficult and complex.

We want safety solutions to have withstood the test of time when we finally deploy in the critical moments of AGI development. We don't want to be using cutting edge methods that haven't undergone extensive testing to be what we finally use to try and solve alignment. Doing so exposes us to a high risk of unforeseen failures that a long period of testing and development may have uncovered.

Early safety work because seems important, time-sensitive, tractable, and informative.

The importance of AI safety work is outlined in [Why is safety important for smarter-than-human AI?](#). We see the problem as time-sensitive as a result of:

- [neglectedness](#) — Only a handful of people are currently working on the open problems outlined in the MIRI technical agenda.

- *apparent difficulty* — Solving the alignment problem may demand a large number of researcher hours, and may also be harder to parallelize than capabilities research.
- *risk asymmetry* — Working on safety too late has larger risks than working on it too early.
- *AI timeline uncertainty* — AI could progress faster than we expect, making it prudent to err on the side of caution.
- *discontinuous progress in AI* — Progress in AI is likely to speed up as we approach general AI. This means that even if AI is many decades away, it would be hazardous to wait for clear signs that general AI is near: clear signs may only arise when it's too late to begin safety work.

We also think it is possible to do useful work in AI safety today, even if smarter-than-human AI is 50 or 100 years away. We think this for a few reasons:

- *lack of basic theory* — If we had simple idealized models of what we mean by correct behavior in autonomous agents, but didn't know how to design practical implementations, this might suggest a need for more hands-on work with developed systems. Instead, however, simple models are what we're missing. Basic theory doesn't necessarily require that we have experience with a software system's implementation details, and the same theory can apply to many different implementations.
- *precedents* — Theoretical computer scientists have had repeated success in developing basic theory in the relative absence of practical implementations. (Well-known examples include Claude Shannon, Alan Turing, Andrey Kolmogorov, and Judea Pearl.)
- *early results* — We've made significant advances since prioritizing some of the theoretical questions we're looking at, especially in *decision theory* and *logical uncertainty*. This suggests that there's low-hanging theoretical fruit to be picked.

Finally, we expect progress in AI safety theory to be useful for improving our understanding of robust AI systems, of the available technical options, and of the broader strategic landscape. In particular, *we expect transparency to be necessary for reliable behavior*, and we think there are basic theoretical prerequisites to making autonomous AI systems transparent to human designers and users.

Having the relevant theory in hand may not be strictly necessary for designing smarter-than-human AI systems — highly reliable agents may need to employ very different architectures or cognitive algorithms than the most easily constructed smarter-than-human systems that exhibit unreliable behavior. For that reason, some fairly general theoretical questions may be more relevant to AI safety work than to mainline AI capabilities work. Key advantages to AI safety work's informativeness, then, include:

- *general value of information* — Making AI safety questions clearer and more precise is likely to give insights into what kinds of formal tools will be useful in answering them. Thus we're less likely to spend our time on entirely the wrong lines of research. Investigating technical problems in this area may also help us develop a

better sense for how difficult the AI problem is, and how difficult the AI alignment problem is.

- *requirements for informative testing* — If the system is opaque, then online testing may not give us most of the information that we need to design safer systems. Humans are opaque general reasoners, and studying the brain has been quite useful for designing more effective AI algorithms, but it has been less useful for building systems for verification and validation.
- *requirements for safe testing* — Extracting information from an opaque system may not be safe, since any sandbox we build may have flaws that are obvious to a superintelligence but not to a human.

Deprecated duplicate which might be worth incorporating

☰ AIs aren't as smart as rats, let alone humans. Isn't it far too early to be worrying about...