Information security considerations for AI and the long term future

New technologies under development, most notably artificial general intelligence (AGI), could pose an existential threat to humanity. We expect significant competitive pressure around the development of AGI, including a significant amount of interest from state actors.

Lennart Heim

May 2, 2022

security

Summary

This post is authored by Jeffrey Ladish, who works on the security team at Anthropic, and Lennart Heim, who works on AI Governance with GovAI (more about us at the end). The views in the post are our own and do not speak for Anthropic or GovAI. This post follows up on Claire Zabel and Luke Muehlhauser's 2019 post, Information security careers for GCR reduction.

We'd like to provide a brief overview on:

1. How information security might impact the long term future
2. Why we'd like the community to prioritize information security

In a following post, we will explore:

1. How you could orient your career toward working on security

Tl;dr:

- New technologies under development, most notably artificial general intelligence (AGI), could pose an existential threat to humanity. We expect significant competitive pressure around the development of AGI, including a significant amount of interest from state actors. As such, there is a large risk that advanced threat actors will hack organizations — that either develop AGI, provide critical supplies to AGI companies, or possess strategically relevant information— to gain a competitive edge in AGI development. Limiting the ability of advanced threat actors to compromise organizations working on AGI development and their suppliers could reduce

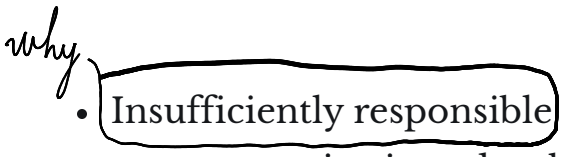
existential risk by decreasing competitive pressures for AGI orgs and making it harder for incautious or uncooperative actors to develop AGI systems.

What is the relevance of information security to the long term future?

The bulk of existential risk likely stems from technologies humans can develop. Among candidate technologies, we think that AGI, and to a lesser extent biotechnology, are most likely to cause human extinction. Among technologies that pose an existential threat, AGI is unique in that it has the potential to permanently shift the risk landscape and enable a stable future without significant risks of extinction or other permanent disasters. While experts in the field have significant disagreements about how to navigate the path to powerful aligned AGI responsibly, they tend to agree that actors that seek to develop AGI should be extremely cautious in the development, testing, and deployment process, given the failures could result in catastrophic risks, including human extinction.

NIST defines information security as "The protection of information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction."

We believe that safe paths to aligned AGI will require extraordinary information security effort for the following reasons:



- Insufficiently responsible or malicious actors are likely to target organizations developing AGI, software and hardware suppliers, and supporting organizations to gain a competitive advantage.
- Thus, protecting those systems will reduce the risk that powerful AGI systems are developed by incautious actors hacking other groups.

Okay, but why is an e*xtraordinary effort* required?

- Plausible paths to AGI, especially if they look like existing AI systems, **expose a *huge amount of attack surface*** because they're built using complex computing systems with very expansive software and hardware supply chains.
- Securing systems as complex as AGI systems is extremely difficult, and **most attempts to do this have failed in the past**, even when the stakes have been large, for example, the Manhattan Project.
- The difficulty of defending a system depends on the threat model, namely **the resources an attacker brings to bear to target a system**. Organizations developing AGI are likely to be targeted by advanced state actors who are amongst the most capable hackers.

Even though this is a challenging problem requiring *extraordinary effort*, we think the investments are worth pursuing, and the **current measures are insufficient**.

What does it mean to protect an AI system?

It can be helpful to imagine concrete scenarios when thinking about AI security — the security of AI systems. Gwern recently wrote a compelling fictional take on one such scenario, including some plausible infosec failures, which we recommend checking out.

Infosec people often think about systems in terms of attack surface. How complex is the system, how many components does it have, how attackable is each component, etc? Developing a modern AI system like GPT-3 involves a lot of researchers and developers -- the GPT-3 paper had 31 authors! Each developer has their own laptop, and needs some amount of access to the models they work on. Usually, code runs on cloud infrastructure, and there are many components that make up that infrastructure. Data needs to be collected and cleaned. Researchers need systems for training and testing models. In the case of GPT-3, an API is created to grant limited access for people outside the company.

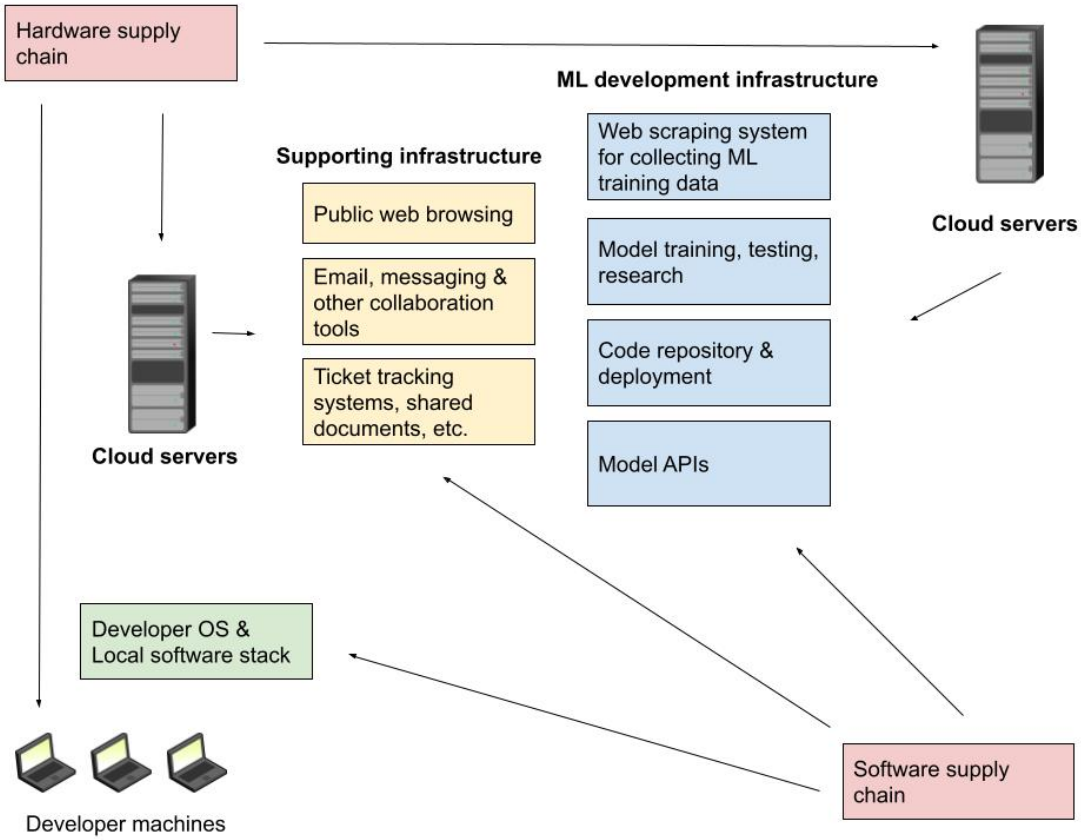


Figure 1: Overview of the active components in the development of an ML system. Each introduces more complexity, expands the threat model, and introduces more potential vulnerabilities.

Most of the components described here contain a staggering amount of complexity (see Figure 1). For example, the Linux kernel alone contains over 20 million lines of code. Each piece of hardware and software is a component that could be exploited. The developer could be using a malicious browser plugin that steals source code, or their antivirus software could be silently exfiltrating critical information. The underlying cloud infrastructure could be compromised, either because of underlying exploits in the cloud provider or because of misconfigurations or vulnerable software introduced by the organization.

And these are just the technical components. What's not

depicted are the humans creating and using the software and hardware systems. Human operators are generally the weakest part of a computing system. For example, developers or system administrators could be phished, phones used for multifactor authentication systems could be SIM-swapped, or passwords for key accounts could be reset by exploiting help desk employees. This is generally described as *social engineering*. In short, modern AI systems are very complex systems built by many people, and thus are fundamentally difficult to secure.

Threat Actors and AGI

Securing a modern AI project against attacks from general cybercriminals — such as ransomware operators, extortionists, etc —is difficult but not extraordinarily difficult. Securing a system as complex as a modern AI system against a state actor or an Advanced Persistent Threat (APT) actor is extremely difficult, as the examples in the reference class failures section demonstrate. We believe it is quite likely that state actors will increasingly target organizations developing AGI for the reasons listed below:

- Near-term AI capabilities may give states significant military advantages
- State-of-the-art AI labs have valuable IP and are thus a rich target for state-sponsored industrial espionage
- Concepts of the strategic value of AGI are present and accessible in our current culture

The US, China, Russia, Israel, and many other states are currently acting as if modern and near-term AI capabilities have the potential to improve strategic weapons systems, including unmanned aerial vehicles (UAVs), unmanned underwater vehicles (UUVs), submarine detection, missile detection systems, and hacking tools.

Some of these areas have already been the subject of successful state hacking activities. Open Philanthropy Project researcher Luke Muelhauser compiles several examples in this document, including: "In 2011-2013, Chinese hackers targeted more than 100 U.S. drone companies, including major defense contractors, and stole designs and other information related to drone technology… Pasternack (2013); Ray (2017)."

It's not evident that espionage aimed at strategic AI technologies will necessarily target labs working on AGI if AGI itself is not recognized as a strategic technology, perhaps because it is perceived as too far out to be beneficial. However, it seems likely that AGI labs will develop more useful applications as they get closer to capable AGI systems, and that some of these applications will touch on areas states care about. For example, China used automated propaganda to target Taiwan elections in 2018 and could plausibly be interested in stealing language models like GPT-3 for this purpose. Code generation technologies like Codex and AlphaCode might have military applications as well, perhaps as hacking tools.

In addition to the direct military utility of AI systems, state-sponsored espionage is also likely to occur for the purpose of economic competition. AI companies have raised $216.6B in

investment, with at least a couple of billion raised by companies specifically trying to pursue AGI development which makes them a valuable economic target. As already outlined, stealing IP for AGI development itself or required critical resources is of interest to malicious actors, such as nations or companies. If IP violations are hard to detect or enforce, it makes industrial espionage especially attractive.

The most concerning reason that states may target AGI labs is that they may believe that AGI is a strategically important technology. Militaries and intelligence agencies have access to the same arguments that convinced the EA community that AI risk is important. Nick Bostrom discusses the potential strategic implications of AGI in Superintelligence (2014, ch. 5). Like with the early development of nuclear weapons, states may fear that their rivals might develop AGI before them.

Note that it is the perception of state actors which matters in these cases. States may perceive new AI capabilities as having strategic benefits even if they do not. Likewise, if AI technologies have the potential for strategic impact but state actors do not recognize this, then state espionage is less of a concern.

Other targets

Next to targeting organizations that develop AGI systems, we also think that organizations that either (a) supply critical resources to AGI labs or (b) research and develop AGI governance strategies are also at risk:

- Suppliers of critical resources, such as compute infrastructure or software, are of significant relevance for AGI organizations. These supplies are often coming from external vendors. Targeting vendors is a common practice for state actors trying to gain access to high value targets. In addition, stealing intellectual property (IP) from external vendors could boost a malicious actor's capabilities of developing relevant resources (which it might have previously been excluded from due to export bans).
- We think that AI Governance researchers may also be targets if they are perceived to possess strategic information about AI systems in development. State actors are likely especially interested in obtaining intelligence on national organizations and those relevant to the military.

Reference Class Failures

In trying to understand the difficulty of securing AI systems, it is useful to look at notable failures in securing critical information and ways to mitigate them. There is no shortage of examples — for a more comprehensive list, see this shared list of examples of high-stake information security breaches by Luke Muehlhauser. We want to present some recent examples — walking on the spectrum from highly technical state-sponsored attacks to relatively simple but devastating social engineering attacks.

One recent example of a highly technical attack is the Pegasus 0-click exploit. Developed by the NSO group, a software firm that sells “technology to combat terror and crime to governments

only", this attack enabled actors to gain full control of the victim's iPhone (reading messages, files, eavesdropping, etc.). It was used to spy on human rights activists and was also connected to the murder of Jamal Khashoggi. As outlined in this blogpost by Google's Project Zero, this attack was highly sophisticated and required a significant amount of resources for its development — costing millions of dollars for state actors to purchase.

On the other side of the spectrum, we have the Twitter account hijacking in 2020, where hackers gained access to internal Twitter tools by manipulating a small number of employees to gain access to their credentials. This then allowed them to take over Twitter accounts of prominent figures, such as Elon Musk and Bill Gates — and all of this was probably done by some teenagers.

Another more recent attack, which also likely relied heavily on social engineering tactics, is the hack of NVIDIA by the Lapsus$ group. This attack is especially interesting as NVIDIA is an important actor in developing AI chips. Their intellectual property (IP) being leaked and potentially being used by less cautious actors could accelerate competitors' efforts in developing more powerful AI chips while actively violating others' IP norms. Notably, while the target of the hacking group is a target of interest to state actors, we have some first hints that this attack might actually have been conducted by a number of teenagers.

The Twitter account hijacking and recent NVIDIA hack are notable as the required resources for those attacks were relatively small. More significant efforts and good security practices could have mitigated those attacks or at least made them significantly

more expensive. Companies that are of strategic importance, or in respect to this article, organizations relevant to the development of AGI systems, should benefit from the best security and be on a similar security level, as national governments — given their importance for (national) security.

The applicability of the security mindset to alignment work

> Good engineering involves thinking about how things can be made to work; the security mindset involves thinking about how things can be made to fail.
>
> — Bruce Schneier

In addition to securing information that could play an outsized role in the future of humanity, we think that the information security field also showcases ways of thinking that are essential for AI alignment efforts. These ideas are outlined in Eliezer Yudkowsky's post Security Mindset and Ordinary Paranoia and follow up post Security Mindset and the Logistic Success Curve. The security mindset is the practice of looking at a system through the lens of adversarial optimization, not just looking for ways to exploit the system, but looking for systemic weakness that might be abused even if the path to exploitation is not obvious. In the second post, Eliezer describes the kind of scenario where the security mindset is crucial to building a safe system:

> …this scenario might hold generally wherever we demand robustness of a complex system that is being subjected to strong external or internal optimization pressures… Pressures that strongly promote the probabilities of particular states of affairs via optimization that searches across a large and complex state space…Pressures which therefore in turn subject other subparts of the system to selection for weird states and previously unenvisioned execution paths… Especially if some of these pressures may be in some sense creative and find states of the system or environment that surprise us or violate our surface generalizations…

This scenario describes security critical code like operating system kernels, and even more so describes AGI systems. AGI systems are incredibly dangerous because they are powerful, opaque optimizers and human engineers are unlikely to have a good understanding of the scope, scale, or target of their optimization power as they are being created. We believe this task is likely to fail without a serious application of the security mindset.

Applying the security mindset means going beyond merely imagining ways your system might fail. For example, if you're concerned your code-writing model-with-uncertain-capabilities might constitute a security threat, and you decide that isolating it in a Docker container is sufficient, then you have failed to apply the security mindset. If your interpretability tools keep showing your models say things its internal representation knows are false, and you decide to use the output of the interpretability tools to train the model to avoid falsehoods, then you have failed to apply the security mindset.

The security mindset is hard to learn and harder to teach. Still, we think that more interplay between the information security community and the AI alignment community could help more AI researchers build skills in this area and improve the security culture of AI organizations.

Building the relevant information security infrastructure now

We think that labs developing powerful AGI systems should prioritize building secure information systems to protect against the theft and abuse of these systems. In addition, we also think that other relevant actors, such as organizations working on strategic research or critical suppliers, are increasingly becoming a target and also need to invest in information security controls.

Our appeal to the AI alignment and governance community is to take information security seriously now, in order to build a firm foundation as threats intensify. Security is not a feature that's easy to tack on later. There is ample low-hanging fruit — using up to date devices and software, end-to-end encryption, strong multi-factor authentication, etc. Setting up these controls is an excellent first step and high investment return. However, robust information security controls will require investment and difficult tradeoffs. People are usually the weak point of information systems. Therefore, training and background checks are essential.

Information security needs are likely to become much more demanding as AGI labs, and those associated with AGI labs, are targeted by increasingly persistent and sophisticated attacks. In the recent Lapsus$ attacks, personal accounts were often targeted to gain access to 2FA systems to compromise company accounts. Tools like Pegasus, already utilized by agencies in dozens of countries, could easily be leveraged against those working on AI policy and research.

While we could make many more specific security recommendations, we want to emphasize the importance of threat awareness and investment in secure infrastructure, rather than any specific control. That being said, if you do have questions about specific security controls, or want help making your org more secure, feel free to reach out! Part of our goal is to help organizations just starting to think about security get started, as well as helping existing organizations to ramp up their security programs.

Conclusion

In this post, we've presented our argument for the importance of information security for the long term future. In the next post, we'll give some concrete suggestions for ways people could contribute to the problem, including:

- 1) how to know if an infosec career is a good idea for you
- 2) how to orient your career toward information security and
- 3) how others working to reduce AI risk can acquire and incorporate infosec skills into their existing work

In the meantime, you can engage with others on related discussions in the Information Security in Effective Altruism Facebook group.

About us

About Jeffrey

- I work on the security team at Anthropic (we're hiring!), and am also working on AI security field building and strategy. I've worn several hats in my infosec career: I worked as a security engineer at Concur Technologies, led security efforts for the cryptocurrency company, Reserve, and then started my own security consultancy business. I also spent a couple of years exploring existential and catastrophic risks from nuclear weapons and biotechnology. You can find some of my work here: https://jeffreyladish.com

About Lennart

- I work on the intersection of AI hardware and AI Governance. Before that, I studied Computer Engineering and have a long-standing interest in security (though I never worked professionally full-time in this field). I used to work on security as a research assistant in wireless and networked systems or in my leisure time, mostly on embedded systems and webservers.

Acknowledgements

Many thanks to Ben Mann, Luke Muehlhauser, Markus Anderljung, and Leah McCuan for feedback on this post.

← Previous Next →

{myname} [ät] heim.xyz