Security and AI alignment

Paul Christiano Follow
Oct 14, 2016 · 8 min read

I am interested in the *alignment* problem: building powerful AI systems so that they are trying to do what we want them to do.

I don't have as much intrinsic interest in the *security* problem, of protecting AI systems from adversaries who want to manipulate their behavior.

That said, I'm starting to feel that working at the intersection of AI and security may be a good way to make progress on alignment, and that many problems in alignment might be naturally understood and approached as security problems. This post probably won't say much new, but I wanted to explain why I think this.

My view on this topic was substantially influenced by Ian Goodfellow at OpenAI (who expects that if we are able to solve security problems then we will also be able to solve alignment problems) and to a lesser extent Peter Eckersley at EFF.

In section 1 I give some examples of problems that I see as both security and alignment problems. In fact I think that *most* alignment problems will first arise as security problems.

In section 2 I give a grab bag of considerations that make me more enthusiastic about security.

In section 3 I discuss some security problems that I don't think are as interesting from an alignment perspective.

1. Examples

Generating adversarial inputs. I am concerned about powerful AI systems operating correctly during training and then failing catastrophically at test time, either due to a change in context or to encountering rare failure-inducing inputs. I think that these situations will eventually cause a lot of trouble when they arise organically, but I think the cleanest examples are currently and will continue to be due to adversaries.

Suppose that I have a household robot with a camera and flexible actuators, that learns how to directly map visual data to actions using contemporary techniques. Our experience with adversarial examples—see especially this—suggest that an adversary could drop off a digital display on my porch that, if viewed by the robot, would cause it to execute a destructive sequence of actions—perhaps escalating the attacker's access, ultimately compromising the robot and stealing my stuff. (Ian pointed out the near-term plausibility of this kind of attack to me, and similar ideas make periodic appearances in futurist folklore.)

Exploiting misalignment. I am concerned about powerful AI systems maximizing objectives that don't exactly match their users' preferences. In the long run I think that this will be a serious problem even in the absence of adversaries. But in the short run, I think that subtle misalignment will provide opportunities for an attacker to cause meaningful damage, and that avoiding these problems will be much harder than avoiding autogenous risks—and so will force us to adopt robust and principled solutions rather than relying on a patchwork of ad hoc solutions that may break down if AI improves quickly.

For example, suppose that an agent is purchasing items on my behalf optimizing for my satisfaction. Even if the agent uses a very sophisticated measurement of user satisfaction, and even if the agent itself has limited opportunities to corrupt the evaluation process, other people have strong motives to corrupt the evaluation process. If the AI is capable of sophisticated planning, or if some corrupted evaluations make it into training data, then a competent AI system might effectively collude with would-be attackers in order to achieve a high satisfaction rating.

Extracting information (covert malicious behavior). I am concerned about AI systems behaving in ways that are damaging yet invisible even to a sophisticated and aligned reward function. In the very long run I think this problem will arise naturally, but in the short term it can already appear in the context of privacy and data exfiltration.

If an AI system has access to some data, then it is very difficult to certify that its behavior does not leak that data. Such leakage is a great example of a behavior which is undesirable but extremely difficult to disincentivize in an RL framework. In fact, the possibility of steganography shows that this problem is inherently asymmetrical,

such that there is no general way to look at some behavior and determine if it leaks.

Realistically we can't hope to avoid leaking *any* information, but we might hope to avoid *gratuitously* leaking information. Note that this is different from the more typical privacy concern, which is that an algorithm may leak information about training data. Here we are concerned with algorithms that take observable actions in the world based on private information, such that we are forced to leak some information, but want to avoid leaking unnecessarily much.

For example, suppose that I am training an RL agent to act in a financial market, and that this algorithm is given access to some proprietary information in order to make better trades. We don't want to leak the private information any faster than is necessary to make good trades. But an adversary can potentially extract all of this information, if they can (1) get access to some of it (for example because they can replicate some of our research, or because they can discover the data with a time delay), and (2) have the ability to "give" money to our agent (for example by making carefully-chosen bad trades). They could then could train the algorithm to leak the proprietary data on demand (for example by changing low-order bits of its bids after observing an appropriate signal). This would be very hard to discourage using conventional RL unless we have some insight into what our agent is actually doing/thinking, because there is no way that the reward function can tell if data was leaked.

2. Motivation

- An AI system that **really works** should continue to work even in the presence of adversaries, as long as we maintain the integrity of the AI's code/hardware/rewards/etc. At some point we might give up on building AI's that work robustly in the presence of adversaries, but I don't think we need to get pessimistic yet.

- If our AI systems fail in certain "hard" situations, those situations are likely to be engineered by an adversary before they arise naturally. So alignment problems will probably first appear as security vulnerabilities.

- Summarizing, the last two points: if something is an alignment problem then it's a security problem, and if something is a security problem then it's an alignment problem. (Though see the caveats in section 3 below.)

- Security problems will typically sound more realistic than analogous alignment problems, precisely because they will happen sooner. For example, the cascading collapse scenario in this post or the steganography example in this post may look outlandish as autogenous failures but are plausible security concerns. The treacherous turn sounds less improbable when it just involves an AI changing the permissions on a file so that an external adversary can compromise the machine.

- If our goal is building robust systems, asking "how could an adversary make this fail?" seems way better than asking "has this failed yet?" or "can we see an obvious reason it will fail soon?" This is inspired by MIRI's interest in the security mindset.

- It is often not clear when we should be worried about a gap between our system's behavior and the "ideal" behavior. I think "can it be exploited by an adversary?" is a surprisingly good test that more people should use, while also being intuitive and sounding sensible to a broad audience.

- I think we should aim for robustness to adversarial errors. This assumption is unusual, and today it is hard to justify to AI researchers. But I think it is much more natural from a security perspective.

- My goal is ultimately to build AI systems that continue to represent our interests in a world containing powerful adversaries (e.g. unaligned AI systems which are competing for resources). Coping with powerful adversaries seems closely aligned with the traditional outlook in security. (Theorists also work with adversarial inputs and Byzantine failures, but they are happiest when the problem is really well-defined.)

- There are going to be lots of high-profile security problems long before we have powerful AI systems. I think many of these failures will be usefully analogous to possible catastrophic AI alignment failures. If we are searching for contemporary analogies to our long-term concerns about AI, security is probably the place to look.

3. Non-examples

I am excited about improving *the security of AI systems themselves*; I think that security issues *related to AI* tend to be less interesting from an alignment perspective.

Many security problems don't have direct relevance to alignment:

- Reliance on AI systems will increase the importance of conventional security problems. For example, an adversary who corrupts the machine on which important AI systems are being trained might eventually be able to cause physical damage or steal money on a massive scale. I think that these AI-adjacent security problems will become increasingly important over time, but I don't see them as directly relevant to alignment.

- AI may change the security landscape by accelerating automation of offense/defense/monitoring. Again, this seems quite important but does not seem directly relevant to alignment, **except** insofar as security is a domain where we are especially likely to apply AI in the presence of adversaries, which highlights other important security issues.

- AI systems often pool large amounts of training data, allowing an attacker to influence the system's behavior by corrupting that data. I think this is an important problem and I have worked on it. But in the context of alignment, I think we should assume that we have access to a secure source of labels. That said, our systems may still learn from very large amounts of insecure unlabelled data, and I think that corruption of unlabelled data is a really important attack vector to consider (along with adversarial inputs, both at test time and training time).

Other problems do have relevance to both security and alignment, but look likely to be less impactful than work on the security of AI systems themselves:

- We could try to isolate AI systems, preventing them from having unintended influences on the outside world or learning unintended data about the outside world. This project has some relevance to alignment, but I don't currently consider it to be an especially promising direction.

- Formal verification is a tool for building secure systems. As described above, this may become more important in a world with AI, but does not seem directly relevant to alignment. Formal verification can *also* be used to design components of AI systems that definitely satisfy some formal contract. Figuring out what kind of contract we might want to satisfy seems relevant to alignment; but I'm less excited about improving our ability to formally verify that we've satisfied some particular contract.

Conclusion

I think that long-term concerns about alignment provide a compelling reason to try harder to connect work on alignment to foreseeable security concerns, to increase investment in understanding the security of AI systems, and to help researchers in alignment and security have a clearer picture of the other field's goals and priorities.

Perhaps the lowest-hanging fruit is for researchers interested in AI alignment to more often use security examples, and to not shy away from it when people respond "that sounds like a security problem." Hopefully this will have the dual benefits of making exotic alignment problems sound less speculative, and helping build bridges between people with interests in alignment and security.