Two guarantees

Paul Christiano Follow
Apr 9, 2018 · 6 min read

When I imagine proving something about an AI, or making an inductive argument about amplification, I think about two different guarantees:

1. The AI behaves well *on average* over some particular distribution of inputs. (The performance guarantee.)
2. The AI never does anything "actively malicious," on *any* input. (The control guarantee.)

Definitions

We can define "behaves well on average" by looking at the average impact of decisions on our utility function. This post addresses some subtleties with that definition, including inconsistent preferences, different scales of impact, and long-term consequences. Actually evaluating this quantity, so that it can be used in training, requires amplification and informed oversight.

My intuition about "actively malicious" is best captured by the (informal) idea of incorrigible optimization: our AI should never be actively trying to undermine our understanding or control of the situation. This property seems "easy," in the sense that it can be satisfied without any domain knowledge or expertise, which makes me optimistic that it is possible to satisfy it for every input.

Rationale

Rationale for the performance guarantee

The performance guarantee is what does the work; it's the reason you'd want an AI instead of a pet rock.

If we construct the reward so it reflects the difference in utility between different actions, then the average case is actually a pretty good thing to optimize.

In reality the data distribution is constantly changing over time. So we want an online guarantee, maximizing the average case performance over a varying sequence, rather than a batch guarantee. Normally deep learning is designed for the batch setting under the iid assumption, and data distributions change slowly enough that this is acceptable. Overall, I don't think this distinction is a huge deal.

Note that the "average case" guarantee still reflects the overseer's desire for corrigibility. This is formally reflected when (a) the overseer evaluates the expected utility impact of a decision by marginalizing over all of the mistakes that the AI might make in the future, (b) the overseer evaluates individual decisions by assigning a large penalty for incorrigibly doing the wrong thing, but only a small penalty for behaving corrigibly but slightly inefficiently.

Rationale for control guarantee

Unfortunately, it is only possible to get an average performance guarantee when the rewards are small relative to the total amount of data. If a few bad decisions can destroy all of the value from a lifetime of good decisions, then traditional ML can't do well on average.

If an AI is *trying* to cause a catastrophe, it can do so very quickly even in apparently benign situations (by breaking the mechanisms that humans use to remain in control of the situation). The goal of the control guarantee is to prevent this kind of catastrophe.

Similarly, if an AI is trying to be difficult, it could intentionally undermine whatever testing regime we use (by failing selectively on inputs that don't look like tests). The control guarantee should also rule out this kind of intentional sabotage.

Hopefully the control guarantee cuts off the worst catastrophes, making the performance guarantee meaningful.

Other catastrophes

Even if our AI satisfies the control guarantee, it can make other kinds of catastrophic mistakes (just as an aligned human could make a catastrophic mistake). For example, the system might be in charge of running a nuclear defense system, and a single failure could lead to a nuclear war. Or an attacker might fool the AI into executing an instruction by impersonating the operator, which in turn open the system to further attack.

Neither the control nor performance guarantee directly address these problems. Instead, anyone who deploys an AI needs to be aware of the system's limits, to test the system to see where it might fail, to design mechanisms with redundancy, to protect the system for attackers, and to avoid incorrectly assuming perfect performance. (The same measures we would take if delegating to a human who sometimes made mistakes.)

Corrigibility

The performance and control guarantees interact to create a system that is corrigible: the performance guarantee ensures the system is typically trying to give the human more effective understanding and control over the situation. The control guarantee ensures the system isn't undermining those measures, for example by constructing a fake narrative for the human or illusory control, by leaving backdoors that can be exploited, etc.

The performance guarantee leaves open the possibility that the system will sometimes fail to inform or empower the human effectively. But as long as those are failures aren't optimized to be unrecoverable, it seems the human can work around them by recognizing the shortcoming and having the AI optimize for the robustness of human control and understanding.

I think there is a meaningful analogy between this picture and this post about monitoring and whistleblowing. and I have a vague intuition that there is some important underlying dynamic that could be better understood.

Amplification

Amplification and performance

Amplification solves a task by (adaptively) breaking it into several subtasks, solving the subtasks, and combining the results.

If we are making an inductive argument about amplification, then the performance guarantee implies the expected *average* performance on subtasks is good, and requires that the expected performance on the overall task is good.

If some of the subtasks make "honest mistakes," then I think this is a reasonable hope.

In order to make it work we definitely need to have an unequal weighting over subtasks, which corresponds to an unequal sampling of subtasks during training. For example, if we are trying to figure out if any of {A, B, C} are positive, we proceed by independently estimating A, B, C, and we conclude that B is positive, then the estimation of B was a more important subtask than the estimation of A or C.

Even with arbitrary weights on subtasks, preserving average performance might seem too optimistic. In the previous example, if we conclude that all of A, B, C are negative, then any one of the three errors could lead to an erroneous overall conclusion. So if our original error rate was 5%, then our new error rate will be 15%.

In cases like this, it's wrong of think of "error rate" as the measure of performance that should be preserved inductively. Some questions are harder than others, and should be expected to have higher error rate—if any rational agent tries to figure out whether any one of N quantities is positive, without any clever trick for figuring it out, then they'll have a harder time than if they were just trying to estimate whether a single quantity is positive. This makes it harder to talk precisely about the guarantees of amplification, but leaves me feeling relatively optimistic about eventually finding a working analysis.

Amplification and control

I find the above intuition plausible when errors are "honest mistakes." If instead one of the subtasks is optimized to cause trouble, then I think it is very hard (and likely impossible) to correct the problem. So in order to have any hope of maintaining performance, we need to guarantee that none of the subtask solutions are actively malicious.

But in order to guarantee that a subtask isn't actively malicious, we also need to assume that none of *its* subtasks are actively malicious. If we want to carry out the induction to a reasonable depth, this basically requires a worst-case guarantee. Otherwise even a single failure, in one leaf of an exponentially large tree, could cause trouble.

Unfortunately, a human probably does not satisfy the control guarantee, since they may behave very strangely on some small subset of possible inputs or with small probability. I originally described security amplification and reliability amplification to gradually eliminate the human control failures (as well as new

failures that crop up, e.g. from memetic selection inside the amplification process).

Interestingly, some ways of achieving the control guarantee in the distillation step could also be applied even if the overseer *doesn't* satisfy the control guarantee, and so could be used as a replacement for security/reliability amplification. This isn't as crazy as it sounds, since the ability to look at the entire code of a learned model gives us a substantial advantage over dealing with an opaque overseer. It gives a further reason to prioritize techniques for worst-case performance (and particularly interpretability).

Conclusion

We can't guarantee an AI is aligned if we have only an average-case guarantee, *or* only a worst-case guarantee. So achieving both seems like the "minimum viable product" for alignment research.

My original intuition (in mid-2016) was that having two separate guarantees must be an ugly hack, and that the real goal should encapsulate both. That's no longer so clear to me: I think these two properties interact surprisingly nicely, such that they may actually suffice to get good behavior even though it looks like a weird combination. At the same time, I think attempts to capture both are much less promising than I'd initially believed.

I still think we need more clarity about what we should be trying to prove. I think that having two separate guarantees, one for the worst case and one for the average case, is the current best guess and is the most promising starting point for further research. In the short term, my focus will be on understanding ways in which this structure is inadequate and on independently refining each of these two subgoals.