Counterfactual oversight vs. training data

Paul Christiano Follow
Oct 3, 2015 · 5 min read

I have written a lot recently about counterfactual human oversight. This idea has made me much more optimistic about AI control proposals based on supervised learning. But counterfactual oversight looks superficially kind of weird, and unlike the kind of thing that would appear in practice. I think that is mostly a presentation issue, and that in fact analysis of counterfactual oversight applies more or less directly to the kinds of supervised learning systems that people are likely to build.

Status quo

Consider the normal workflow for training a supervised learning system:

- Collect and label training data.
- Train a learning system.
- Deploy that system.

This workflow can run into a few well-known problems:

- The problem is not stationary, and over time the training data becomes less relevant.
- There are spurious correlations in the training data that don’t generalize to the test data.

How do we address these problems?

- Continue to gather training data while the system is deployed. Periodically adjust the learned model (or maybe use an online approach if the setting calls for it).
- Try to make the training data as similar as possible to the test data.

Ideally, our training data would be a random subset of the test data, and and we would train continuously. Of course, if data needs labelling and the task is performed frequently, it will be impractical to label it all by hand. So instead we might label a small fraction of it.

In most cases these measures aren't necessary if we are mindful of the possible problems, and we will instead address problems in the cheapest way available.

Counterfactual oversight

Counterfactual oversight consists of labelling a random subset of data and using it as online training data. The key difference is that any given data point *may* become a training data point, with the decision made *after* the learning system has made a decision about it. As long as the randomization is unpredictable to the learner, this gives us a formal guarantee that there can't be any noticeable difference between the training and test data. And therefore if our learner behaves well on training data, it really must behave well on test data.

In most cases, I expect that this solution is overkill, so if it's expensive we can probably do something cheaper. For example, I would be surprised if researchers ever really needed to use a secure cryptographic RNG to decide what cases to include in the training set.

But when thinking about scalability to extreme cases, it seems worthwhile to check that there is a (relatively) cheap solution which is essentially perfectly robust. This then gives us a nice model to use when thinking about those extreme cases, and if we notice any problems with the extreme solution we can pay extra attention to them. If the extreme solution works and isn't too expensive, we can be reassured that people will find *some* solution that works and isn't too expensive, whether or not it's the particular one we imagined.

When describing counterfactual oversight I also usually imagine that our algorithms can deal with sequential data rather than discarding information about time. This allows them to adjust to trends in the data, or to anticipate changes based on available information, rather than anchoring their behavior to past experiences. I suspect this will be an important issue in the future, but for the most part it can be ignored for now. This ability isn't a requirement for applying counterfactual oversight—the point is that counterfactual oversight allows us to apply this ability while retaining strong formal guarantees.

A final difference is one of language. I describe counterfactual oversight as the learner "trying to do what the evaluator would rate highly," with the evaluator picking some random data points to actually rate. This would more traditionally be described as the learner trying to solve the underlying problem, with the evaluator providing some useful training data to help. I think that this difference in language results from a difference of perspective—I am trying to pay close attention to what exactly the system is actually doing. This vantage point seems especially suitable for thinking about AI control, but it doesn't directly translate to any technical differences in the systems being discussed. I also happen to think that many practicing AI researchers could stand to be a bit more precise in this respect, though they probably shouldn't go as far as I do and it's not so important one way or the other.

So why think about it?

If counterfactual oversight is very similar to existing practice, why bother thinking about it?

I am interested in understanding in advance what issues will arise as we try to scale existing approaches to AI control to very powerful systems—to whatever extent that is possible.

The supervised learning paradigm has a number of distinctive challenges, for example based on non-stationary data, the possibility of spurious correlations in training data, and the availability and cost of supervision. Today these problems are real but typically manageable; it's conceivable that they will become more severe as learning systems become more powerful. It's natural to ask whether they are likely to become *much* more severe, and in particular whether they call into question supervised learning as a paradigm for controlling very powerful AI systems.

Counterfactual oversight seems to address many of these problems in a robust way, suggesting that they won't be deal-breakers "in the limit." For example, a human-level online learner under counterfactual oversight is unlikely to predictably behave badly because of spurious correlations in the training data. This suggests that such spurious correlations are unlikely to be deal-breakers. Similarly, it seems that the amount of data required for a sophisticated semi-supervised learner using counterfactual oversight would be comparable to the amount of data needed by a completely unsupervised learner, suggesting that the availability of training data is not a deal-breaker either.

Prior to considering counterfactual oversight, I had expected that training processes involving humans were unlikely to be suitable as part of the *definition* of correct behavior for sophisticated AI systems —that they could only be used to help provide auxiliary information that would be useful to an AI in achieving goals defined by some other means. Thinking through the consequences of counterfactual oversight has largely addressed the narrow versions of this concern (though there are many closely related issues that remain open).

As far as I can tell, practicing AI researchers mostly didn't have this concern, which is fine. I think there is room for some people to approach long-term issues with a more theoretical and cautious stance; I'm pretty optimistic that the problems with scalability that seem especially challenging in theory are also likely to generate interesting practical questions for AI researchers today.