Human-in-the-counterfactual-loop

Paul Christiano Follow
Jan 20, 2015 · 7 min read

Consider an autonomous system which is buying or selling assets, operating heavy machinery, or making war. People often discuss three kinds of supervision:

- **Fully autonomous.** The system operates without human supervision.
- **Human-in-the-loop.** The system does what a human tells it to.
- **Human-on-the-loop.** A human oversees the system and may override its behavior if necessary.

In the context of human-level AI, I think a fourth model of supervision is interesting:

- **Human-in-the-counterfactual-loop**. Each time the system wants to act, it consults a human with a very small probability. The system does what it thinks a human *would have* told it to do if the human had been consulted.

I think this model isolates an interesting characteristic of autonomy, and is likely to be a useful form of supervision for powerful systems.

Why bother with prediction?

Why would you have a human in the counterfactual loop rather than the actual loop?

Machines can make lots of decisions quickly. One drawback of human involvement is that human time is expensive and human responses are high-latency.

If you are involved with each decision with small probability, then:

- You can oversee many more decisions, much more cheaply. One person can oversee many machines acting in parallel.

- The system might act much faster than you come to a conclusion. You could take minutes or hours to reach a conclusion that the machine made in seconds. (This is particularly important when one person is overseeing many machines, since they may need to spend some time understanding the context in which a decision was made.)
- We can reduce human error by allowing human overseers to think for longer about each decision.

Why bother being in the loop?

Human oversight seems particularly important when it is difficult to explicitly state all of the goals of a system. For example, we might be concerned that any simple rule for asset purchases might sometimes misfire. If you buy any asset whose expected price is higher tomorrow than today, the asset might be frozen, or the purchase might have ethical repercussions, or the transaction might be politically or legally problematic, or so on. Unanticipated possibilities become more extreme and more worrying as we broaden the space of possible actions or make higher-stakes decisions.

With a human in the counterfactual loop, our system applies its predictive abilities to notice conditions that we would have objected to, if we had foreseen them. This partially addresses many of these concerns.

Although I'll describe such a system as having a "human" in the loop, the role could instead be played by any group of humans. As an extreme example, we could have a democracy-in-the-loop (though there would be a few practical issues). This is possible because the turnaround time in a counterfactual loop can be so much longer than in a normal loop.

More details

Every time an autonomous system with a human in the counterfactual loop needs human feedback, it first flips a biased coin which comes up heads with small probability (say 0.001%). If the coin comes up heads, then the system consults a human, supplying them with relevant context and getting their feedback (though the actual feedback can be provided long after the decision). If the coin comes up tails, the system does not consult a human. In either case, it predicts what feedback the human would eventually provide if the coin came up heads, and uses the predicted feedback as an input into

its decision-making process (as if the human were actually in the loop and had provided that feedback).

This requires an algorithm that can effectively predict human judgments. The more complex the domain, the harder this prediction problem is. There is no guarantee that we have predictors powerful enough to help oversee the autonomous systems that we build. But the predictors never need to exercise superhuman ability, and so we will certainly have good enough predictors by the time we reach broadly human-level AI.

The predictor needs to be able to make counterfactual predictions. But assuming that the predictor can't predict which way the coin will turn up, this amounts to a conditional prediction: *given* that the coined turned up heads, what do I predict the human will tell me? Any useful predicting system must be able to make predictions of this form.

Why actually consult the overseer?

We actually consult the overseer in order to provide a concrete experiment for the predictor to predict. If the predictor is able to consider more extreme counterfactuals, then we may not need to even have a human around. We could instead say "Predict what Bob *would* say, if we had trained him to oversee you." This doesn't actually change the details much. But for this post, I'll conservatively assume that we need to actually set up the experiment in order to get good predictions about it.

There are two forces that keep us from reducing the probability of human involvement to 0:

Keeping us sharp. If the human only steps in once every century, the human may have forgotten what they are supposed to be doing. But this is a *very* weak requirement, and it applies at the level of the human operator rather than of the machine. So one human could potentially oversee a warehouse of thousands of machines, carefully reviewing only a few decisions each day (or month).

Training data. A sophisticated AI could predict the human overseer's behavior using very few direct observations, by leveraging other related information. For example, if I hear someone say "Please be careful with those boxes" I will expect them to object to any plan that involves dropping the boxes. I don't need a library of training data which includes boxes being dropped. That said, this still requires

some contact or common understanding between the overseer and the machines being overseen, so that they can understand how their decisions will be evaluated.

For a weak system, actually consulting the human overseer regularly may be the only way to make good predictions about the overseer. This data can still be shared across many similar systems, potentially driving the per-system cost very low.

Many systems will be somewhere in between these extremes, able to leverage some general information (whether from observations of the overseer, or general heuristics installed by the designers of the system) but would need to get some direct observations of the overseer in order to make good predictions.

Discussion

A human in the counterfactual loop eliminates some but not all of the concerns with automation (and so serves as a thought experiment to identify which aspects of human control are most important). In particular, such a system will only take actions *it thinks* the operator is OK with—but it might be wrong about what the operator thinks is OK.

Strong predictors

For an AI with a very good understanding of humans, prediction errors would be rare, and so a human in the counterfactual loop may be better than the real thing (since they can think longer).

There are also advantages over approaches without counterfactual oversight:

- The predicted response of the human can change as the environment changes (assuming that the predictor knows about the changes, and can predict how the human would respond to them).
- The human overseer can handle all of the factors that we care about, which may otherwise be difficult to formalize.
- Because the human only needs to be able to learn about a decision when prompted, they need not have extensive training in advance. So putting a predicted human in the loop may be a lot cheaper than spending engineer time making sure you have correctly defined a domain-specific decision criterion.

With a strong predictor, any failure of the system can be traced to either (1) a failure of the predictor, or (2) a failure that also would have occurred with human oversight.

Weak predictors

For an AI with a weak understanding of humans, this proposal separates *human judgment* from *human preferences*. A system with a human in the counterfactual loop may make errors that would be caught by a human overseer, but it won't fail by virtue of having different preferences than a human overseer.

That said, in some contexts a weak predictor might do much better than a human—especially when decisions are made quickly or under emotional pressure (in the counterfactual loop, the human can take a long time and consider the issue carefully). In some cases this could be tested directly: we can record the context surrounding a decision, and empirically evaluate whether an AI or human is better able to predict the results of an extended deliberative process.

If this test suggests that the AI is a better predictor of deliberative human judgment, then the objection to automated decision-making becomes substantially weaker.

A digression on weapons. Autonomous weapons are not my primary concern in this post, but it's an example that would be hard to avoid. I think that a human in the counterfactual loop would be a legitimate response to the most common concerns with human rights violations stemming from autonomous weapons—assuming the system passed the test described in the last paragraph (which may already be possible in simple contexts). However, I'm afraid that these humans rights issues obscure a more serious concern with autonomous weapons, namely that they may radically reduce the expense and risk of ending human lives.

Robustness

All approaches to oversight require the overseen system to "go along with it": if a flexible autonomous system wanted to cut the human out of the loop, they would have many opportunities to do so. Any successful approach to oversight must either make this impossible or (more likely) ensure that the autonomous system has no motivation to cut the human out of the loop.

Having a human in the counterfactual loop doesn't fundamentally change these dynamics. It introduces a new point of failure for the

human's involvement (the predictor which is responsible for anticipating the human's response). But given how many possible points of failure there already are (the communication channel, the human's understanding of the situation, the part of the code that responds to the human's judgment, etc.), this is a minor addition.

Using a counterfactual loop makes it more realistic for the human to provide fine-grained oversight, improving the probability that they can oversee the detailed plans made by the machine or the full range of actions that might cut them out of the loop.

Conclusion

For powerful AI systems, effective human oversight may be cheaper than it appears; we should keep this in mind when thinking about superintelligent AI. The idea of counterfactual human oversight may also be a useful building block more generally.