Unsupervised learning and AI control

Paul Christiano Follow
Nov 30, 2015 · 6 min read

Reinforcement learning systems optimize for an objective defined by external feedback—anything from successful prediction of image labels to good performance in a game. (I am using "reinforcement" learning very broadly, including e.g. supervised learning.) I think it's safe to say that reinforcement/supervised learning is overwhelmingly dominant in machine learning today.

I think it is also safe to say that many researchers think this will eventually change, even for the domains and techniques where supervised learning is currently most dominant—e.g. Yann LeCun wrote last year "everyone [in deep learning] agrees that the future is in unsupervised learning," while acknowledging "the recent practical success of deep learning in image and speech all use purely supervised backprop."

Improved unsupervised learning might be good news for AI control, and the prospect of unsupervised learning plays a major role in informal discussions of AI control. If unsupervised learning can extract robust and meaningful concepts, these concepts may be available to communicate goals and preferences to AI systems, who can then pursue what we actually care about rather than an approximation defined by external feedback.

My take

(*See also:* Caveats *below.)*

I think that we should try to address the AI control problem using reinforcement learning, rather than assuming that the problem will be made much easier by progress in unsupervised learning.

More specifically: I think that we should assume that we can train systems to be good at specific tasks for which we can provide feedback on performance, but beyond that we should not rely on strong assumptions about their internal representations or other characteristics of their behavior.

By “broadly construed” I mean to include capabilities like efficient prediction, semi-supervised learning, and density estimation. These areas capture many *capabilities* that people have in mind when they discuss unsupervised learning. But in terms of their relevance to AI control, they are quite similar to supervised learning, and they certainly fit within the more specific framework in the last paragraph.

In contrast, researchers often express the hope that sophisticated AI systems will discover many of the same robust concepts that humans use when they reason about the world, and that these concepts will be sufficiently precise, and in the right format, and sufficiently aligned with the human concepts, that they can be used directly to issue commands or specify goals.

My recommendation is to treat future capabilities as being similar-in-kind to contemporary reinforcement learning, though applying in broader domains and with more efficient use of data, rather than making optimistic and somewhat vague assumptions about what unsupervised learning will do for us.

Justification

My reasons for focusing on reinforcement/supervised learning are:

- I don’t think that we have much idea what future unsupervised learning will look like, and I think that specific assumptions about it are likely to be wrong. These assumptions are often very imprecise and intuitive, which I feel makes them especially suspect and hard to reason about. I think that our best *single* model of future unsupervised learning may well be current unsupervised learning, which is pretty well-modeled as good prediction and semi-supervised learning.
- I think there is a good case that research in AI control should focus on existing techniques. We can understand these techniques much better than unknown future techniques; we can do empirical work on these techniques; and this work will be especially relevant if AI control happens to become important surprisingly soon. This argument suggests we should work with reinforcement learning while that’s what we have, and think more about unsupervised learning as the techniques mature.
- Even if unsupervised learning is the dominant paradigm in the future, it seems quite plausible that control techniques based on reinforcement learning will remain relevant. For example, reinforcement learning may still remain one useful technique

amongst many, and may play a significant role in the internal organization of powerful AI systems.
As a special case, many possible "unsupervised" learning strategies involve interacting reinforcement systems with cleverly designed objectives. Other strategies use unsupervised learning to build useful representations, but only extract useful behavior from these representations by supervised fine-tuning.

Caveats

To clarify:

- I think we will make significant advances in unsupervised learning, and these advances will be relevant to AI control. I am making a methodological claim about how to think effectively about and do useful research on AI control right now—not a strong prediction about the future of machine learning.
- I think that the prospect of improved unsupervised learning should make us more optimistic about AI control. I don't think that modest changes in our overall level of optimism will have big effects on what we should do, unless we end up *very* optimistic (which I don't think is justified), or we started out *very* pessimistic (which I don't think is justified either).
- A better understanding of unsupervised learning may be helpful for AI control. At face value this doesn't look like a promising project for a researcher concerned with AI control (since understanding unsupervised learning is a hot topic), but it still has some positive differential impact and it might end up being a winner. And of course I am happy to be supportive of AI researchers who are working on differentially useful projects, even if they aren't maximally differentially useful from my perspective.
- I assume that future systems will (eventually) be able to make superhumanly efficient use of data. I object to making more detailed assumptions about the features learned by unsupervised learning, *not* to the assumption that we will eventually realize the practical goals of unsupervised learning. For example, I think that we should expect AI systems to develop conceptual understanding that allows them to quickly learn the meaning of a new word or even do zero-shot learning.
- I think it is worthwhile to pay attention to continuing progress in unsupervised learning and to adjust our approach to AI control

as we learn more and can make firmer predictions about what will be possible.

An example of the distinction

Suppose that we train a learner to recognize which scenes contain humans.

In order to analyze how the behavior of this system will scale, I would think in detail about the training process, the objective it is optimizing, and what behaviors would optimize that objective. For example, if the system is trained to reproduce human labels, then I expect more sophisticated systems to converge to the human's labels.

We might hope that in the future an unsupervised approach would learn to identify which scenes "really" contained humans, and could make correct judgments even in cases where the human would err or in domains where we couldn't actually elicit a human label.

I would recommend against making this kind assumption until we have learned more about unsupervised learning.

This discussion becomes more complex and important when we think about messier concepts like "good" or "what Hugh wants." It is relatively clear what the reinforcement model predicts—powerful AI systems will make increasingly accurate predictions about how a human labeler would label the given data. It's not clear exactly what the unsupervised approach would do, and I think that we shouldn't count on it doing something that is "good."

The current situation

Very few people have thought seriously about how to handle AI control if reinforcement learning remains as dominant as it currently is; I think that few people are optimistic enough to think that the task is possible and pessimistic enough to think that it may be necessary.

When I've discussed the issue with AI researchers, they seem to have strong expectations that progress in unsupervised learning will obviate many of the concerns with AI control for reinforcement learning (and with AI control more broadly), allowing users to e.g. provide natural language instructions that will be correctly understood and implemented. I think this is a very reasonable hypothesis, and that the views of AI researchers are an important source of evidence for it.

But I'm not yet persuaded that this is much more likely than not. I also don't think that most AI researchers have considered the question in much detail or engaged with substantive arguments that these problems are real. I'm not even sure that *they* think that this optimistic hypothesis is much more likely than not, rather than considering it the dominant hypothesis or most promising approach.

MIRI researchers mostly seem to take the opposite view—that unsupervised learning definitely won't address these problems by default. But the MIRI research agenda responds by focusing on a number of problems they see as needed to make unsupervised learning work for goal specification—ontology identification, multi-level world models, ambiguity identification and operator modeling. I think most MIRI researchers feel that we are probably doomed if we are stuck with the kind of reinforcement learning that is available today. (This is my unconfirmed impression based on informal discussions.)

Upshot

Unsupervised learning will probably improve significantly before AI control becomes a serious problem. Nevertheless, I think that researchers interested in AI control should focus on handling reinforcement learning systems of the kind that already exist.