Ambitious vs. narrow value learning

by **Paul Christiano** 11th Jan 2019

Value Learning | Frontpage

(Re)Posted as part of the AI Alignment Forum sequence on Value Learning°.

> ***Rohin's note:*** *The definition of narrow value learning in the previous post focused on the fact that the resulting behavior is limited to some domain. The definition in this post focuses on learning instrumental goals and values. While the definitions are different, I have used the same term for both because I believe that they are both pointing at the same underlying concept. (I do not know if Paul agrees.) I'm including this post to give a different perspective on what I mean by narrow value learning, before delving into conceptual ideas within narrow value learning.*

Suppose I'm trying to build an AI system that "learns what I want" and helps me get it. I think that people sometimes use different interpretations of this goal. At two extremes of a spectrum of possible interpretations:

- The AI learns my preferences over (very) long-term outcomes. If I were to die tomorrow, it could continue pursuing my goals without me; if humanity were to disappear tomorrow, it could rebuild the kind of civilization we would want; *etc.* The AI might pursue radically different subgoals than I would on the scale of months and years, if it thinks that those subgoals better achieve what I really want.
- The AI learns the narrower subgoals and instrumental values I am pursuing. It learns that I am trying to schedule an appointment for Tuesday and that I want to avoid inconveniencing anyone, or that I am trying to fix a particular bug without introducing new problems, *etc.* It does not make any effort to pursue wildly different short-term goals than I would in order to better realize my long-term values, though it may help me correct some errors that I would be able to recognize as such.

I think that many researchers interested in AI safety per se mostly think about the former. I think that researchers with a more practical orientation mostly think about the latter.

The ambitious approach

The maximally ambitious approach has a natural theoretical appeal, but it also seems quite hard. It requires understanding human preferences in domains where humans are typically very uncertain, and where our answers to simple questions are often inconsistent, like how we should balance our own welfare with the welfare of others, or what kinds of activities we really want to pursue vs. enjoying in the moment. (It seems unlikely to me that there is a unified notion of "what I want" in many of these cases.) It also requires extrapolation to radically unfamiliar domains, where we will need to make decisions about issues like population ethics, what kinds of creatures do we care about, and unforeseen new technologies.

I have written about this problem, pointing out that it is unclear how you would solve it even with an unlimited amount of computing power. My impression is that most practitioners don't think of this problem even as a long-term research goal—it's a qualitatively different project without direct relevance to the kinds of problems they want to solve.

The narrow approach

The narrow approach looks relatively tractable and well-motivated by existing problems. We want to build machines that helps us do the things we want to do, and to that end they need to be able to understand what we are trying to do and what instrumental values guide our behavior. To the extent that our "preferences" are underdetermined or inconsistent, we are happy if our systems at least do as well as a human, and make the kinds of improvements that humans would reliably consider improvements.

But it's not clear that anything short of the maximally ambitious approach can solve the problem we ultimately care about. A sufficiently clever machine will be able to make long-term plans that are significantly better than human plans. In the long run, we will want to be able to use AI abilities to make these improved plans, and to generally perform tasks in ways that humans would never think of perform them—going far beyond correcting simple errors that can be easily recognized as such.

In defense of the narrow approach

I think that the narrow approach probably takes us much further than it at first appears. I've written about these arguments before, which are for the most part similar to the reasons that approval-directed agents or directly mimicking human behavior might work, but I'll quickly summarize them again:

Instrumental goals

Humans have many clear instrumental goals like "remaining in effective control of the AI systems I deploy," "acquiring resources and other influence in the world," or "better understanding the world and what I want." A value learner may able to learn robust preferences like these and pursue those instrumental goals using all of its ingenuity. Such AI's would not necessarily be at a significant disadvantage with respect to normal competition, yet the resources they acquired would remain under meaningful human control (if that's what their users would prefer).

This requires learning robust formulations of concepts like "meaningful control," but it does not require making inferences about cases where humans have conflicting intuitions, nor considering cases which are radically different from those encountered in training—AI systems can continue to gather training data and query their users even as the nature of human-AI interactions changes (if that's what their users would prefer).

Process

Even if we can't infer human preferences over very distant objects, we might be able to infer human preferences well enough to guide a process of deliberation (real or hypothetical). Using the inferred preferences of the human could help eliminate some of the errors that a human would traditionally make during deliberation. Presumably these errors run counter to a deliberator's short-term objectives, if those objectives are properly understood, and this judgment doesn't require a direct understanding of the deliberator's big-picture values.

This kind of error-correction could be used as a complement to other kinds of idealization, like providing the human a lot of time, allowing them to consult a large community of advisors, or allowing them to use automated tools.

Such a process of error-corrected deliberation could itself be used to provide a more robust definition of values or a more forward looking criterion of action, such as "an outcome/action is valuable to the extent that I would/did judge it valuable after extensive deliberation."

Bootstrapping

By interacting with AI assistants, humans can potentially form and execute very sophisticated plans; if so, simply helping them achieve their short-term goals may be all that is needed. For some discussion of this idea, see these three posts.

Conclusion

I think that researchers interested in scalable AI control have been too quick to dismiss "narrow" value learning as unrelated to their core challenge. Overall I expect that the availability of effective narrow value learning would significantly simplify the AI control problem even for superintelligent systems, though at the moment we don't understand the relationship very well.

(Thanks to Andreas Stuhlmüller and Owain Evans for helpful discussion.)

• • •

This was originally posted here on 4th October, 2015.

Value Learning 9 Frontpage

Previous:

What is narrow value learning?

3 comments 23 karma

Next:

Human-AI Interaction

No comments 34 karma

Log in to save where you left off

Mentioned in

46 Four visions of Transformative AI success
21 Conclusion to the sequence on value learning
17 [Intro to brain-like-AGI safety] 10. The alignment problem
9 Inferring utility functions from locally non-transitive preferences
16 New year, new research agenda post
Load More (5/7)

15 comments, sorted by top scoring

[−] **Wei Dai** 5y 4 0

> Even if we can't infer human preferences over very distant objects, we might be able to infer human preferences well enough to guide a process of deliberation (real or hypothetical). Using the inferred preferences of the human could help eliminate some of the errors that a human would traditionally make during deliberation.

This assumes (depends on) that human deliberation is good/safe because humans have good preferences about deliberation. But what if human deliberation is only good/safe because of the constraints that we face? Example of what I mean: Someone wants to self-modify to have 100% certainty that God exists before doing further deliberation, but can't, and as a result eventually realizes through deliberation that having 100% certainty that God exists is actually not a good idea.

This can be considered an instance of the more general concern I have about humans not being safe°, especially under distributional shift.

ETA: Here° is another example from our own community:

> Meanwhile, a few years ago when I first learned about the concept of updatelessness, I resolved to be updateless from that point onwards. I am now glad that I couldn't actually commit to anything then.

[−] **Wei Dai** 5y 1 0

How would this kind of narrow value learning work in a mathematical or algorithmic sense? For example, one question I have is, since instrumental goals and values can be invalidated by environmental changes (e.g., I'd stop valuing US dollars if I couldn't buy things with them anymore), how does the value learner know when that has happened? Are there any papers I can read or tutorials I can watch to learn about this? Or feel free to give an explanation here if it's simple enough.

[−] **Rohin Shah** 5y 2 0

> How would this kind of narrow value learning work in a mathematical or algorithmic sense?

I'm not sure I understand the question. Inverse reinforcement learning, preference learning (eg. deep RL from human preferences) and inverse reward design are some existing examples of narrow value learning.

> since instrumental goals and values can be invalidated by environmental changes (e.g., I'd stop valuing US dollars if I couldn't buy things with them anymore), how does the value learner know when that has happened?

By default, it doesn't. You have to put active work to make sure the value learner continues to do what you want. Afaik there isn't any literature on this.

[−] **Wei Dai** 5y 1 0

> I'm not sure I understand the question. Inverse reinforcement learning, preference learning (eg. deep RL from human preferences) and inverse reward design are some existing examples of narrow value learning.

Thanks for the existing examples, which are helpful, but I guess what I was trying to ask was, is there a mathematical theory of instrumental value learning, that we can expect practical algorithms to better approximate over time, which would let us predict what future algorithms might look like or be able to do?

> You have to put active work to make sure the value learner continues to do what you want.

"You" meaning the user? Does the user need to know when they need to provide the AI with more training data? (For example, if there was a massive devaluation of the US dollar, they need to predict that the AI might sell all their other possessions for dollars, and actively provide the AI with more training data before that happens?) Or can we expect the AI to know when it should ask the user for more training data? If the latter, what can we expect the AI to do in the meantime (e.g., if the user is asleep and it can't ask)?

[−] **Rohin Shah** 5y 1 0

> is there a mathematical theory of instrumental value learning, that we can expect practical algorithms to better approximate over time, which would let us predict what future algorithms might look like or be able to do?

Not to my knowledge, though partly I'm hoping that this sequence will encourage more work on that front. Eg. I'd be interested in analyzing a variant of CIRL where the human's reward exogenously changes over time. This is clearly an incorrect model of what actually happens, and in particular breaks down once the AI system can predict how the human's reward will change over time, but I expect there to be interesting insights to be gained from a conceptual analysis.

> "You" meaning the user?

Yes.

> Does the user need to know when they need to provide the AI with more training data? Or can we expect the AI to know when it should ask the user for more training data?

Hopefully not, I meant only that the user would need to provide more data, it seems quite possible to have the AI system figure out when that is necessary.

> If the latter, what can we expect the AI to do in the meantime (e.g., if the user is asleep and it can't ask)?

I don't imagine this as "suddenly the reward changed dramatically and following the old reward is catastrophic", more like "the human's priorities have shifted slightly, you need to account for this at *some* point or you'll get compounding errors, but it's not crucial that you do it immediately". To answer your question more directly, in the meantime the AI can continue doing what it was doing in the past (and in cases where it is unsure, it preserves option value, though one would hope this doesn't need to be explicitly coded in and arises from "try to help the human").

[−] **Wei Dai** 5y 1 0

> and in cases where it is unsure, it preserves option value, though one would hope this doesn't need to be explicitly coded in and arises from "try to help the human"

Do you mean something like, the AI is learning instrumental values, option value is a kind of instrumental value, so hopefully the AI can learn to preserve option value? If so, I worry that option value may be a particularly complex type of instrumental value that would be hard to learn and hard to generalize well, so the AI wouldn't be able to correctly preserve option value in cases where it is unsure. It may seem simple to us only because option value is simple *given* a set of terminal goals, but the narrow value learner wouldn't know those terminal goals. Kind of like how a big multiplication table is simple and easy to generalize if you knew that everything is connected by the concept of multiplication, but complex and hard to generalize if you learn it as a series of brute facts.

This is the type of question that I'd want a theory of instrumental value learning to address.

[−] **Rohin Shah** 5y 1 0

Certainly in the case where you are uncertain about long-term terminal goals, you should realize that you want to preserve option value.

> Do you mean something like, the AI is learning instrumental values, option value is a kind of instrumental value, so hopefully the AI can learn to preserve option value?

That is not what I meant, I meant something more like "if you are trying to help someone and you are unsure of what they want, preserving option value is a robustly good thing to do".

I could imagine that this doesn't happen with instrumental goals because with a short enough time horizon, it could be better in expectation to bet on the most likely goal and pursue that.

> This is the type of question that I'd want a theory of instrumental value learning to address.

I do intend to look into questions about time horizons, option value, risk aversion, etc., probably over the summer. I'm not sure I'd classify it as a "theory of instrumental value learning" but it should be relevant to the questions we're talking about here.

[−] **Wei Dai** 5y 2 0

> Certainly in the case where you are uncertain about long-term terminal goals, you should realize that you want to preserve option value.

This is confusing because aren't we talking about a narrow value learner which isn't even trying to learn long-term terminal goals? How would it realize that it wants to preserve option value? Is the idea that if it tries to learn instrumental goals that are long-term and have uncertainty about those, that would be enough for it to want to preserve option value? But if it can do that, why can't it just try to learn terminal goals? What is it doing that's different from a value learner that's trying to learn terminal goals?

> That is not what I meant, I meant something more like "if you are trying to help someone and you are unsure of what they want, preserving option value is a robustly good thing to do".

But the AI does not have an intuition notion of "help someone" that they can use. Since we've been talking about narrow value learning, I'm assuming the AI just has an algorithm that does some form of narrow value learning.

> I do intend to look into questions about time horizons, option value, risk aversion, etc., probably over the summer. I'm not sure I'd classify it as a "theory of instrumental value learning" but it should be relevant to the questions we're talking about here.

That sounds very useful, but I'm not sure it would be enough to resolve my confusions around narrow value learning. But it might so we can certainly come back to these questions after you do that.

[−] **Rohin Shah** 5y 1 0

> This is confusing because aren't we talking about a narrow value learner which isn't even trying to learn long-term terminal goals?

Sorry, I meant that in the sense "this works if you have long-term goals, it plausibly could also work when you have instrumental goals".

> Is the idea that if it tries to learn instrumental goals that are long-term and have uncertainty about those, that would be enough for it to want to preserve option value?

This seems plausible, though I don't want to make that claim yet. For example, an instrumental goal I have is to acquire resources such that I have influence over the future, which is a long-term goal.

> But if it can do that, why can't it just try to learn terminal goals? What is it doing that's different from a value learner that's trying to learn terminal goals?

There seem to be philosophical difficulties with trying to learn *all* of your terminal goals *exactly*. But I'm not opposed to it also trying to learn my terminal goals, as long as it can account for the fact that I don't know my terminal goals yet. (Whereas I do know some of my long-term instrumental goals.)

> But the AI does not have an intuition notion of "help someone" that they can use. Since we've been talking about narrow value learning, I'm assuming the AI just has an algorithm that does some form of narrow value learning.

Yeah, I was using "help someone" as a shorthand for "optimize for their goals as determined by narrow value learning".

Btw, you should take most of this as speculation about what could happen with narrow value learning systems, not as claimed fact. I'm hypothesizing that narrow value learning systems could be working with sufficiently long-term instrumental goals that under uncertainty preserving option value arises as a good strategy. I don't think that this is inevitable, and expect that the actual answer depends strongly on the details of how the narrow value learning is done.

Wei Dai 5y 1 0

> For example, an instrumental goal I have is to acquire resources such that I have influence over the future, which is a long-term goal.

This is a bit tangential but I've noticed that this instrumental goal can be interpreted in a narrow way or a broader/more ambitious way. For example if the AI just learns a list of useful resources I instrumentally value and how many utils per gram (or whatever suitable unit of measurement) each resource is worth under various kinds of circumstances it has seen in the past, that would be relatively easy to learn but is not going to generalize well. Or alternatively the AI could learn "resources" as a general concept, and be able to infer what counts as resources in a new environment it hasn't seen before and how different resources should be traded off against each other as circumstances change. This would be (putting aside human safety problems) really useful in a general and robust way but it's unclear to me that we can realistically hope for such a thing. (For example is it possible to determine the the relative values of different resources in a novel situation if you don't at least have a rough idea what they'll ultimately be used for?)

Do you see the former as the end goal of narrow value learning, or the latter? When you talk about narrow value learning AI preserving option value, do you have the former or the latter kind of narrow value learning in mind?

BTW what is the success story for narrow value learning? Is it the same as for norm-following AI °? Is the success story for all of the approaches described in this sequence essentially #4 in this list °?

> I don't think that this is inevitable, and expect that the actual answer depends strongly on the details of how the narrow value learning is done.

Understood, thanks for the clarification.

Rohin Shah 5y 2 0

> Do you see the former as the end goal of narrow value learning, or the latter? When you talk about narrow value learning AI preserving option value, do you have the former or the latter kind of narrow value learning in mind?

Closer to the latter, i.e. the broad/ambitious kind. I think it's a generally reasonable model to imagine human-like narrow value learning capabilities. When I talk about option value, I also have the broad version in mind.

> BTW what is the success story for narrow value learning? Is it the same as for norm-following AI °? Is the success story for all of the approaches described in this sequence essentially #4 in this list °?

Yeah, it's either #4 or #5 (it remains to be seen whether narrow value learning is sufficient for #5, but it seems possible to me).

It's also plausible to me that the success story *for today's research* is "research on narrow value learning illuminates a lot of important properties that we would want our AI systems to have, which influences AGI research positively". That is, as a result of the research on narrow value learning, the AGI systems we build

will be better at narrow value learning than they otherwise would have been, even though they aren't using a specific narrow value learning algorithm that we develop. And this in turn leads to success story #4 or #5.

This also feels like it illustrates a difference in our thinking -- you seem to be thinking about *explicit* techniques for alignment, whereas I'm often thinking of *implicit* ones. It might also be something more like outside view vs. inside view (where somewhat paradoxically I'm on the outside view side)? I'm currently far too confused about this distinction to explain it well, but I wanted to flag it as a thing causing confusion in our dialogue that I hope to resolve. (And I would write a post about it if I did.)

Wei Dai 5y 2 0

> I think it's a generally reasonable model to imagine human-like narrow value learning capabilities.

I think I'm much more skeptical about this. Humans generally have a fairly good idea of other humans' "terminal values" and their narrow value learning is strongly informed by that. I don't see how the more ambitious kind of narrow value learning could work without this knowledge. As I wrote in the previous comment, "For example is it possible to determine the the relative values of different resources in a novel situation if you don't at least have a rough idea what they'll ultimately be used for?"

Maybe you're imagining that the AI has learned an equally good idea of humans' "terminal values" but they're just being used to help with narrow value learning instead of being maximized directly, similar to how a human assistant doesn't try to directly maximize their boss's terminal values? So essentially "narrow value learning" is like an explicit algorithmic implementation of corrigibility (instead of learning corrigibility from humans like in IDA). Is this a correct view of what you have in mind?

> Yeah, it's either #4 or #5 (it remains to be seen whether narrow value learning is sufficient for #5, but it seems possible to me).

I guess there's also hope that it could be used in some hybrid approach ° to help achieve any of the other positive outcomes.

> That is, as a result of the research on narrow value learning, the AGI systems we build will be better at narrow value learning than they otherwise would have been, even though they aren't using a specific narrow value learning algorithm that we develop.

Do you have an example of something like this happening in the past that could help me understand what you mean here?

Rohin Shah 5y 1 0

> I think I'm much more skeptical about this. Humans generally have a fairly good idea of other humans' "terminal values" and their narrow value learning is strongly informed by that. I don't see how the more ambitious kind of narrow value learning could work without this knowledge. As I wrote in the previous comment, "For example is it possible to determine the the relative values of different resources in a novel situation if you don't at least have a rough idea what they'll ultimately be used for?"

> Maybe you're imagining that the AI has learned an equally good idea of humans' "terminal values" but they're just being used to help with narrow value learning instead of being maximized directly, similar to how a human assistant doesn't try to directly maximize their boss's terminal values? So essentially "narrow value learning" is like an explicit algorithmic implementation of corrigibility (instead of learning corrigibility from humans like in IDA). Is this a correct view of what you have in mind?

Partly I want to claim "explicit vs implicit" and table it for now.

But yes, I am expecting that the AI has learned some idea of "terminal values" that helps with learning narrow values, eg. the AI can at least predict that we personally don't want to die, it seems likely that we want sentience and conscious experience to continue on into the future, we probably want happiness rather than suffering, etc. but still not be able to turn it into a function to be maximized directly.

It seems probably true that most of the hope that I'm expressing here can be thought of as "let's use narrow value learning to create an algorithmic implementation of corrigibility". I feel much better about that description of my position than any other so far, though it still feels slightly wrong in a way I can't put my finger on.

> I guess there's also hope that it could be used in some hybrid approach ° to help achieve any of the other positive outcomes.

Yeah, that seems right. I was describing success stories that could potentially occur with only narrow value learning.

> Do you have an example of something like this happening in the past that could help me understand what you mean here?

The VNM rationality theorem has (probably) helped me be more effective at my goals (eg. by being more willing to maximize expected donation dollars rather than putting a premium on low risk) even though I am not literally running expected utility maximization.

I could believe that the knowledge of Dijkstra's algorithm significantly influenced the design of the Internet (specifically the IP layer), even though the Internet doesn't use it.

Insights from social science about what makes a "good explanation" are influencing interpretability research currently.

Einstein was probably only able to come up with the theory of relativity because he already understood Newton's theory, even though Newton's theory was in some sense wrong.

[−] **Wei Dai** 5y 2 0

Ok, I think I mostly understand now, but it seems like I had to do a lot of guessing and asking questions to figure out what your hopes are for the future of narrow value learning and how you see it potentially fit into the big picture for long term AI safety, which are important motivations for this part of the sequence. Did you write about them somewhere that I missed, or were you planning to write about them later? If later, I think it would have been better to write about them at the same time that you introduced narrow value learning, so readers have some idea of why they should pay attention to it.

(This is mostly feedback for future reference, but I guess you could also add to previous posts for the benefit of future readers.)

[−] **Rohin Shah** 5y 2 0

Yeah, this seems right. I didn't include them because it's a lot more fuzzy and intuition-y than everything else that I've written. (This wasn't an explicit, conscious choice; more like when I generated the list of things I wanted to write about, this wasn't on it because it was insufficiently crystallized.) I agree that it really should be in the sequence somewhere, I'll probably add it to the post on narrow value learning some time after the sequence is done.

AI safety without goal-directed behavior° *very* vaguely gestures in the right direction, but there's no reasonable way for a reader to figure out my hopes for narrow value learning from that post alone.

Moderation Log