Machines learning values*

Steve Petersen

April 27, 2020

Abstract

We would very much like any nascent superintelligence to share our core values—but it seems our values are too complex to program or hardwire explicitly. Our best hope may thus be to design any potential superintelligence to *learn* values like ours. This "value learning" approach to AI safety faces three particularly philosophical problems: first, it is unclear how any intelligent system could learn its final values, since to judge one supposedly "final" value against another seems to require a further background standard for judging. Second, it is unclear how to determine the content of a system's values based on its physical or computational structure. Finally, there is the distinctly ethical question of which values we should best aim for the system to learn. This paper outlines a potential answer to these interrelated problems, centering on a "miktotelic" proposal for blending a complex, learnable final value out of many simpler ones.

Here's a serious problem. Suppose, as many think, that humans will someday be able to create an artificial *superintelligence*—an intelligence whose intellectual capacities outstrip ours the way ours outstrip those of ants. Such a superintelligence is likely to have values quite different from ours; just as we wouldn't expect it to love doughnuts or sunny beaches, so we shouldn't assume it would share our desires for social connection, or high art, or the general welfare. It seems an intelligent system could value *any* goal, no matter how foreign to us; as the standard trope goes, a superintelligence *could* in principle value ever more paperclips in the world. In efficient pursuit of such a foreign value the superintelligence could wipe us out with no more thought or malice than we give to anthills on a construction site.[1]

(I will be taking it for granted that this is a serious worry. If you are one of the many who feel it is easy to dismiss the problem, I can here only urge you to read

*Thanks to Einar Duenger Bøhn, John Danaher, Matthew Liao, Eric Schwitzgebel, Marija Slavkovik, and two anonymous reviewers.

[1]Of course no one thinks the "paperclip maximizer" is likely; it's just to illustrate that without the particularities of human evolutionary history, an AI is free to have *any* goal. To think no intelligence could value such a thing is mere anthropomorphizing—no intelligence we know *today* would value such a thing. The example is originally from Bostrom (2003).

The comparison to our concern for ants is also a common trope in the literature, and goes back at least as far as Daniel Dewey in Andersen (2013).

Nick Bostrom's *Superintelligence* (2014), or some of these other references.[2] I for one went into the literature skeptical, and came out scared.)

A natural solution to this problem is to attempt to design the superintelligence with fundamental values similar enough to ours. This has become known as the goal of *value alignment*. This proposed solution to the superintelligence problem has its own problem, though: human-friendly values are too complex for us to hardwire or program explicitly. After all, as Bostrom points out, philosophers do not even agree on how to paraphrase key values like *happiness* into other similarly abstract terms, let alone into concrete computational primitives (loc. 4332).

A natural solution to the complexity of values problem (for the value alignment solution to the superintelligence problem) is at least as old as Alan Turing, but getting notoriously more successful all the time. When some computational task is too complex to program explicitly, you must design the machine to *learn* to achieve it. This technique has already worked on tasks like winning go games against professional humans and scoring above human average on reading comprehension tests. In this case, we would like to make sure any nascent superintelligence will learn complex, human-friendly values. This constitutes the subfield of *value learning*, in the intersection of machine learning and value alignment.[3]

To many—including me—value learning seems like our best hope for getting non-disastrous superintelligence. But of course, value learning also faces problems. This paper concentrates on three particularly philosophical hurdles for the project. I consider them in order of increasing difficulty; correspondingly, the sections dedicated to them get shorter and sketchier as we go.

Problem one: learning goals in service of another goal is routine for AIs, but in this case we want the potential superintelligence to learn complex "final" values—ends in themselves. But good arguments seem to show *no* cognitive system could learn its final values.

> **Related philosophical issue:** the metaethical debate between *moral rationalism* (according to which, roughly, pure intellect can direct us toward ethical goals) *vs. sentimentalism* (according to which, roughly, reason can have nothing to say about fundamental values).

Problem two: we do not know how to map computational states—especially in connectionist architectures—onto a system's abstract reasoning. In particular,

[2] Chances are very good Bostrom has thoroughly addressed the reasons you are tempted to dismiss the worry. My (2017) paper was the best comfort I could concoct in response to Bostrom, and that comfort was pretty cold. If you don't have time for Bostrom's book, maybe try instead one of these:

- https://www.vox.com/future-perfect/2018/12/21/18126576/ai-artificial-intelligence-machine-learning-safety-alignment
- https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html
- https://futureoflife.org/background/aimyths/

[3] See Soares (2016) or Shah (2018) for an overview.

looking at a machine state is not typically enough to determine the particular content of a system's values. But the particular content is very much at issue in value alignment.

> **Related philosophical issue:** the debate in the philosophy of mind over whether and how *mental content* can be "naturalized"—that is, shown to be a purely physical property (in some broad sense).

Problem three: even if we were perfectly confident of being able to prime the superintelligence to learn any complex values we wanted, there is still the thorny question of *which* values we would like something with amazing superpowers to have.

> **Related philosophical issue:** the traditional philosophical problem of *normative ethics*—the problem of determining what is right and wrong.

I sketch an interrelated solution to these problems, revealed as they are considered in detail. The heart of the proposal is to build a complex, learnable value in a computationally respectable way out of the right blend of simpler values. In the philosophical tradition of resorting to ancient Greek, I call this proposal *miktoteleology* ("blended-goal-studies").

Learning *final* values

The first problem, recall: we want a potential superintelligence to be able to learn a final goal, but there is good reason to think *no* cognitive system can learn a final goal. To see why, it is important first to get clearer on the sense of "value" at play.

It is not clear what exactly it takes for a system to have *real* values. We tend to agree that the system we call "Nick Bostrom" has values, and the system we call "the Great Red Spot of Jupiter" does not. In between are problem cases, like bees, amoebas, and Roombas. For the purposes of saving humanity, we needn't get too hung up on the philosophy here; a superintelligent system that *behaves* in a way consistent with valuing ever more paperclips over anything else is no less dangerous if the philosophers declare on *a priori* grounds that such systems possess no genuine values. Instead, we can be content with what philosophers of mind call the *functionalist* account of mental states, according to which (very roughly) what determines the possession of mental states is the right combinations of system inputs, internal system processing, and system outputs.[4] Broadly speaking, if a system internally processes its sensory input in the right way to generate behavior aimed at maximizing the total number of

[4] For an old but good overview of functionalism, see Churchland (1988). Of course there remain many further interesting philosophical questions about whether such functionalism determines *all* relevant senses of value, meaning, consciousness, ethical worth, and so on. Like many philosophers, I am inclined to say "yes"—but it is beside the point here.

paperclips in the world, then functionalists are happy to say that system does indeed genuinely value a world with more paperclips in the relevant sense.

Now there is a kind of value learning, on this broad functionalist sense, that is relatively straightforward for AIs. For example, the AI AlphaZero was simply taught the rules of chess. After playing itself and learning what works and what doesn't for a few hours, it learned that things tend to go better when you do not give away your queen—it learned to *value* the queen more than the knights (again, and from now on, in our broad functionalist sense). But this kind of value learning is not directly relevant to the value alignment problem. AlphaZero treats the queen as valuable only because it has figured out that typically, the queen helps it achieve its further value of winning chess games. In the odd situation where a queen sacrifice would lead to a win, AlphaZero would happily sacrifice the queen.

Philosophers distinguish *instrumental* values from *final* ones. For AlphaZero, having the queen on the board is usually of instrumental value, because it usually serves as an instrument toward the further goal of winning. But the chess variant of AlphaZero values chess wins "in themselves", not for achieving some further purpose; the wins are of final value for it. For humans, a standard example of an instrumental value is money. We might seek money to be able to afford a vacation, and we might seek a vacation in order to relax, and we might want to relax in order to feel good. If asked why we want to feel good, in turn, we understandably have little to say. The regress of "why" stops at the final goals, which are sought for their own sakes.

It is only learning *final* goals that is philosophically problematic. To see why, consider what is required for a physical system to be able to *learn* something. I assume first that arbitrary changes to a cognitive system do not count as learning; if cosmic rays or a dull hammer rearrange my brain, then even if the resulting cognition is better (no comment), we shouldn't count this improvement as learning. In other words, learning must be purposeful—the result of some cognitive function to adjust other cognitive functions according to feedback. This feedback serves as an internal measure of error, in effect assessing the distance between how things seem to be and how they "should" be. Such error signals thus implicitly contain both a representation of some aspect of the system's current state (how it is now doing) and the goal state (from which it may err). Speaking very loosely, a system with a learning mechanism contains both a "belief" about how the system is doing, and a "desire" for how the system *should* be doing. Speaking more generically and somewhat more strictly, the system has representations with both *indicative* content about how things are (like our beliefs), and *imperative* content about what to do (like our desires).[5]

[5]Beliefs are the paradigmatic indicatives, and desires are the paradigmatic imperatives, but there are surely many levels of mental content that fish or mice or robots might have that are not as sophisticated as beliefs and desires. For a better catalog of ways that our representations differ from those of simpler cognitive systems, see the conclusion of Millikan (1989). (I am using "representation" in a broad sense, roughly synonymous with other philosophical terms of art like "intentionality" and "mental content".)

One helpful approximation is to think of the system's indicatives as afferent information, flowing up from sensory input to report how things are, and the imperatives as efferent information, flowing down toward motor output to bring about helpful actions. Especially given the kind of recurrent feedback between layers in brains, this picture emphasizes that there will not be sharp boundaries between indicatives and imperatives. For example, consider an instrumental goal such as "gather the purple berries." This representation is imperative relative to lower levels of implementation, since it serves as an abstract directive about how to move. But it is also indicative relative to goals like nutrition and survival, since it serves as a hypothesis about how to achieve those further goals. In this sense *instrumental* goals are indicative as well as imperative, and their indicative component makes it easier to see how they can be adjusted and learned when experience intervenes.

Now we are in a better position to see why learning a final goal is problematic. To learn a putatively final goal would be to adjust it based on a measure of success, which is thereby to adjust it against some *further* standard. That would just show the putatively final goal was actually an instrument for the further standard, which is the real final goal. In effect, final goals can have no indicative content, and so no learnable content. Arguments like this, to the effect that reasoning cannot alter final ends, have their roots in Aristotle and David Hume; I have just adapted them somewhat for the context of machine learning, so that we can more clearly see its echoes in the modern debate.[6]

Thus for example Nick Bostrom argues that the standard machine learning technique of reinforcement learning (RL) isn't properly understood as value learning. A reinforcement learner typically gets rewarded for certain kinds of sensory inputs, and uses these reinforcements to update an evaluation function that estimates the expected value of a *policy*—a proposed series of actions (depending on environmental responses), or probability distribution over them.[7] Bostrom points out that "what is being learned" in an RL agent "is not new final values but increasingly accurate estimates of the instrumental values" (loc. 4388). The RL's final value remains its *fixed* reward function.

Bostrom's related concern about using RL agents to learn friendly values is that RL agents are ultimately rewarded by having a certain kind of indicative information stream. This gives any clever such agent incentive to "wirehead"—that is, to hijack its indicative stream to send only maximally rewarding signals. As a simple illustration, imagine a reinforcement learner rewarded for "seeing" (*e.g.*, having information extracted from its cameras contain) gigantic piles of paperclips. A clever such system could simply tape a high-resolution picture of many paperclips in front of its camera, and enjoy constant reward on the cheap. Even better, a truly resourceful system that understood its own design could simply inject the digitization of such an image downstream from its camera,

[6] See Aristotle (350BC) Book III, and Hume (1739) 2.3.3.
[7] The standard RL text is Sutton and Barto (1998).

without any need for the picture or tape.[8] (Thus the term "wireheading", from old experiments using electric current to stimulate mouse brains' reward centers directly.)

Wireheading is just an extreme version of the very human phenomenon of *wishful thinking*, in which we come to believe that things are as we want them to be. More neutrally, wishful thinking involves artificially adjusting the indicative information stream to match the imperative one better. Note if the imperative side is also thoroughly malleable, as it would be in genuine final value learning, there is another potential problem for RL: the learner could instead manipulate the *imperative* stream to match the indicative one. We might call this converse phenomenon *thinkful wishing*, and it too probably occurs in humans—as for example when we decide we didn't really want the grapes that are out of our reach (they are probably sour).[9]

Based on such doubts Bostrom seems to prefer the "utility agent" learning approach from Hibbard (2012) over RL. Utility agents attempt a clean separation between the indicatives and imperatives—roughly a state estimator for the former, and a utility function for the latter. The state estimator tries to figure out which possible world the agent is in (as a probability distribution over them), the utility function scores the worlds for values, and the value learner uses the combination to learn the utility-maximizing policy. Because a paperclip-maximizing utility agent scores a world with more *actual* paperclips higher than a world with mere pictures of paperclips, it would have no reason to pursue a policy designed to bring about the world with mere pictures of paperclips. Everitt and Hutter (2016) point out that "the difference between RL and utility agents is mirrored in the experience machine debate" from Nozick (1974). As they summarize it,

> given the option to enter a machine that will offer you the most pleasant delusions, but make you useless to the "real world", would you enter? An RL agent would enter, but a utility agent would not.[10]

But I suspect the utility agent approach will have similar problems with wishful thinking. As Bostrom is well aware, the ways a world could be are too fine-grained even for a superintelligence to track. (Consider, for starters, all the permutations of particles that would result in a phenomenally identical chair.) This means the utility agent must *abstract* to the relevant aspects of the way the world is—where it seems "relevance" must be determined ultimately by the agent's goals. If the superintelligence is *learning* how best to abstract—as anything worthy of the name must—it must be learning against a standard of success with goals. But here there is danger very like wishful thinking, because

[8] Just in case such short-circuiting sounds at all farfetched, consider that nature designed orgasms to reward reproductive behavior—and that we humans (and many other animals) have found ways to achieve this reward without the intended behavior.

[9] The term "thinkful wishing" is from collaboration with Eric Lormand.

[10] Everitt and Hutter (2016) p. 2, footnote 1. As a reviewer points out, this applies only in general to utility agents; we *could* design ones whose utility function would enjoin them to enter the experience machine.

it is a fine line between learning abstractions in order to better achieve goals efficiently, and learning abstractions to make it look more as though goals were being achieved.[11]

Furthermore, utility agents that are true *value-learners* must be able to adapt their utility functions as well, and this introduces dangers of thinkful wishing in addition to wishful thinking. For example, Bostrom's own favored value-learning utility agent adapts a proposal from Dewey (2011) into what he calls an "AI-VL". Instead of possessing one straightforward utility function, the AI-VL considers a wide range of possible utility functions, and assigns each a weight representing its *guess* that this is the correct utility function, given its estimate of how the world is. (You can imagine the AI-VL implicitly saying, "Given how things appear to me, I am 3% confident that utility function U_1 is the right one, 17% confident it is U_2 instead ... ") In the meantime it treats the weighted average $(.03U_1 + .17U_2 + \ldots)$ as its current utility estimator. You might naturally wonder on what basis the AI-VL could assign or update these guesses about which is the "correct" utility function. The answer is that utility functions are assessed against a background "value criterion."[12]

AI-VL has its problems, of course. For starters, it is "wildly computationally intractable" (loc. 4564). It also pushes much of the problem back a step, into the difficulties of specifying a detailed value criterion that is both largely under our control and computationally inferrable. (The key suggestion later in this paper can be seen as a step toward solving this problem.) Another problem—one more to our point—is that if the system is adjusting its goals based on its estimate of how the world is, there will again be pressure toward thinkful wishing, because its proposed policies are more likely to have higher expected utility if the utility function comes to score easily accessible worlds more highly.[13]

Finally, and even more to our point, the AI-VL still does not seem to *learn* a final goal, because its real final goal seems to be the "value criterion", which assesses utility functions to find the good ones. Bostrom concedes that the value-learning utility agent actually "retains an unchanging final goal", and then says something intriguing:

> Learning does not change the goal. It changes only the AI's beliefs about the goal. (loc. 4473)

If the value-learning superintelligence has a fixed final goal, in what sense is it learning its values? Bostrom suggests here that changing *beliefs about* a fixed

[11] Related ontological concerns are in De Blanc (2011).

[12] Where $U_i(w) \in \mathbb{U}$ is a utility function scoring possible worlds, and $\nu(U_i)$ is the "value criterion" (most generically, "U_i is the correct target utility function"), AI-VL estimates the target utility function and so the value of any possible world as $\hat{U}(w) = \sum_{U_i \in \mathbb{U}} U_i(w) P(\nu(U_i)|w)$.

[13] Everitt and Hutter (2016) propose a value learning system VRL, a hybrid between utility agent and RL, which learns its utility function through reinforcement. Everitt and Hutter then show that a standard such VRL will have incentive to "optimise its evidence" toward "a more easily satisfied utility function" (p. 10)—in other words, to thinkfully wish. They propose a fix for this concern, but rightly worry about its generality.

final goal is sufficient to learn the goal. Note that changing beliefs about a target goal presupposes that the goal starts out sufficiently mysterious to the agent. Bostrom's own example of a value criterion is "maximize the realization of the values [I've] described in [this] envelope." (If we managed to design a superintelligent utility agent trying to learn such a goal, it would have little incentive to harm us along the way, since it would find it fairly probable that harming us would violate the goals written in the envelope.) This illustrates how a utility agent could retain one fixed goal while its particular guesses about the nature of that goal might vary in both content and confidence, as it learns about Bostrom and tries to guess what he might have written.

A more down-to-earth example of a value criterion would be to "do what humans would find most rewarding." Such an agent would have to infer by our behavior—including (defeasible) weight on behavior like our coaching and self-reports—what we would find rewarding. This approach to value learning is called "Inverse Reinforcement Learning", because the agent must learn a reward function from policies and observations rather than, in standard RL, learning a policy from observations and rewards.[14]

Indeed we humans sometimes only learn what's valuable to us after we observe our own behavior—and not necessarily then, either. In other words, *we humans* seem to be final-value learners in this sense, because our own final goals are plausibly quite mysterious to us. Consider for example Ebeneezer Scrooge's transformation in Dickens' *A Christmas Carol* (1843). We might naturally describe his character arc by saying that he used to have the final goal of "hoarding wealth", but through the story's events changed his final value to something like "spreading good cheer" instead. And since this change was not arbitrary, but for the better, we could say he *learned* a new final value.

On the other hand, we might say instead that Scrooge always had the *fixed* but more mysterious goal of "increasing personal happiness", and he changed his beliefs about how best to obtain that one fixed goal. As Aristotle pointed out long ago, "to say that happiness is the chief good seems a platitude, and a clearer account of what it is still desired"[15]—in other words, happiness is one of those opaque, learnable final goals.

Either way, I am happy to say with Bostrom that Scrooge, the inverse reinforcement learner, and the envelope values maximizer are all "learning" new final values in at least this important and relevant sense: they are attempting to *specify* their vague and opaque final goals more precisely. And perhaps it is no coincidence that one of the few ethical views that makes room for reasoning about final ends is called *specificationism*, according to which "at least some practical reasoning consists in filling in overly abstract ends ... to arrive at

[14] See Ng and Russell (2000) for the seminal paper, Sezener (2015) for a more flexible (and more computationally troublesome) take, and Hadfield-Menell et al. (2016) for incorporating the observed agent's feedback ("cooperative inverse reinforcement learning").

[15] Aristotle (350BC), 1097b22.

richer and more concretely specified versions of those ends."[16] So here we have something of a solution to our first value learning problem: how can we *learn* a final value? Answer: if it is abstract enough, we can attempt to *specify* it more concretely.

It may seem obviously unwise to give a potentially superintelligent value learner a deliberately underspecified and mysterious goal. I share this misgiving; I just think providing a precise and unmysterious goal must be even *worse*. For one thing, the danger from superintelligence is not really unpredictability. A monomaniacal superintelligent paperclip maximizer, for example, would be utterly predictable—at least in its final goal—but no less dangerous for that. For another thing, our own values are complex and vague, so we can be confident that a superintelligence with a precise and simply-stated goal (simple enough at least for humans to program it directly) will not align with our interests.[17] After all, if we could specify exactly and briefly what our values consisted in, there would be a lot less moral disagreement in the world.

Another apparent problem with this proposal is its threat of circularity. On this picture, final values can be specified by beliefs; more generally, top-level imperatives can be altered by upstream indicatives. But the indicatives, after all (instrumental goals on down) are aimed ultimately toward fulfilling the top-level imperatives. What, then, is the ultimate arbiter? Or is it possible, as Henry Richardson asks, to do practical deliberation "without an umpire"?[18]

Though problematic, such cases are quotidian. Sometimes, when faced with the tension between a deep *desire* for tasty grapes and a *belief* that they are well out of reach, we keep the desire and alter our instrumental goals, devising new strategies until we come to believe "I can get those grapes" (and eventually "I am tasting yummy grapes"). Other times, the belief that the grapes are unattainable is the relatively stubborn thought, and we attenuate the desire for them instead. Which happens depends on whatever other tiebreakers are nearby in the cognitive system. Philosophers are long familiar with such situations, in which any one element may be revised to satisfy enough of the others, and no elements are needed to be foundational or axiomatic. It comes up in epistemology, for example, where higher-level (more abstract) indicatives conflict with lower-level (more perceptual) ones. Suppose you perceive something truly surprising—perhaps, a tiny flying elephant. In some circumstances you might decide your senses are not currently trustworthy (say, you just took a hallucinogen); in other circumstances you might revise your higher-level beliefs about the probability of such things occurring (say, you are visiting a top-secret genetic engineering lab). In such cases we seek to resolve the conflict while causing the fewest other conflicts and tensions elsewhere. In other words, we seek overall *coherence*. Ethical specificationism suggests we appeal to similar overall coherence considerations when determining

[16] Millgram (2008) p. 744. See also Kolnai (1962), and Richardson (1994) for an extended treatment.

[17] See Yudkowsky (2011).

[18] Richardson (1994) p. 137.

whether the belief should alter the final value (through specification), or the final value should alter the belief (through action to bring about new perceptions).

The exact nature of coherence reasoning is itself a matter needing further specification.[19] The basic idea, though, is to systematize a set of elements between which exist varying degrees of support and tension, typically without holding any special subgroup as inviolable. Thagard and Verbeurgt (1998) and Thagard (1988) suggest that it is best modelled as what computer scientists call a "weighted constraint satisfaction problem."[20] For a simple example, imagine planning the seating chart for a wedding. Between any two guests you might assign some degree of positive or negative conviviality (including perfect neutrality), and then try variations of table assignments to maximize the conviviality total. Optimizing these calculations is in general impossible for even a supercomputer to do in a reasonable amount of time—as anyone who has tried such tasks will be unsurprised to learn.

In our case, seeking coherence among the various and differently-weighted indicatives and imperatives in the system seems to me an especially apt way to capture how abstract content could guide specification of a final goal while not already deductively containing some specification. Since an aim at overall coherence ultimately shapes both the imperatives and indicatives, we *could* say that maximal coherence is the true, final, fixed, unlearnable goal of such an agent—the ultimate "umpire."[21] Indeed, I suspect coherence-seeking is a necessary condition for being an intelligent agent in the first place, and find support in views like that of Friston et al. (2015).

But of course agents could not seek "pure" coherence, for its own sake. The coherence must involve satisfying imperatives already in place for the system, such as for food, or for images of paperclips. We don't want our superintelligence to learn *any* complex, abstract goal. Thus so far we have only the barest hint of high-level design for an agent that can learn complex values: we want it to be a coherence reasoner, able to adjust its final goals (*via* specification) based on its beliefs, while also aiming its beliefs (in particular its assessment of how it's doing) toward satisfaction of (its best current guess at) its final goals. We've already seen two examples of such "coherence" reasoning schemata—inverse reinforcement learning and AI-VL. But how do we engineer a coherence reasoner to learn an abstract, complex, vague goal that *also* has decidedly friendly content? This brings us to our next two problems for value learning.

[19] As Elijah Millgram (2008) puts it, "coherence is a vague concept; we should expect it to require specification" (p. 741). Note in particular that the coherence sought here is not (just) the *probabilistic* coherence demanded by Bayesian reasoning, familiar to many AI theorists.

[20] In collaboration with Millgram, Paul Thagard developed accounts of *deliberative* coherence in Millgram and Thagard (1996) and Thagard and Millgram (1995); see also Thagard (2000). Though inspired by such work, I now lean toward an alternative Millgram also mentions—see *e.g.* Grünwald (2007).

[21] Note Richardson (1994) would not agree; see his section 26. (His account relies instead on a "sovereign deliberator" that I find dubious in light of naturalism and AI.)

Learning *specific* final values

The second philosophical problem implicated in the value alignment problem is to determine the relation between a system's physical or computational structure and that system's values. We have been taking a "functionalist" approach to such questions, where valuing some state roughly means processing observations in a way designed to select actions that achieve that state. But this requires spelling out. Adapting the parable of the thermostat from Daniel Dennett's (1981) paper, we *could* spin functionalist-style stories according to which an ordinary paperclip-manufacturing machine of today "wants" to bend wire into paperclips when it "believes" it is receiving wire in one end, "wants" to sit idle when it "believes" its power is off, and so on. But no one is inclined to say that an ordinary paperclip-making machine of today has a *real* value of making paperclips. Dennett's hypothesis is that we do not attribute making paperclips as a goal to such a machine because it is not very resourceful in achieving it—in other words, on a standard reading of "intelligence" as adaptability in achieving goals, the machine is not *intelligent*. If the wire isn't fed just right or the electricity isn't on, no paperclips will be made.

But now consider variations on ever-more sophisticated and resourceful paperclip-making machines. Suppose it has sensors indicating when it is about to run out of wire, and able to dispatch itself in the direction of more. Suppose it has sensors for, and safeguards against, being turned off or losing a power supply. Suppose it experiments with new paperclip designs, has various ways to sense whether it is successful in making more paperclips, and so on. At some point—at least at the point where it is able to coax us into providing it with more raw materials—the functionalist should say that thing really does, literally, *want* to make paperclips.

This still leaves room for debate over the precise *content* of such values, however—and getting the precise content right is very much at issue in value alignment. Consider a well-worn philosophical illustration of simple but still indeterminate mental content: suppose a small dark patch moving through a frog's visual field causes the frog to snap out its tongue, thereby catching and swallowing a tiny dark metal ball that happened to be sailing by.[22] Between the stimulus and the response, there was some causally-related activity in the frog's brain—the frog was, very broadly speaking, thinking. But what exactly was it thinking *about*? We might naturally say that the frog's brain mistakenly was thinking *hey, a fly*, and so snapped at it. Or perhaps it was just thinking of it more broadly as *insect*? Or more narrowly as a *fly that is nearby and healthy*? Or perhaps, looking up the causal chain for more *distal* causes for the cognition, it was thinking of *food*, or *survival affordance*, or *inclusive genetic fitness enhancer*? Or perhaps we should be looking further *down* the causal chain, to more *proximal* causes—perhaps it was just thinking *hey, a small dark flying thing*, or *hey, a*

[22]The case is discussed extensively in Fodor (1990), but is older than that; the source reference tends to be Lettvin et al. (1959).

spot on the retina. If so, the frog wasn't mistaken at all, since there *was* a small dark flying thing and spot on the retina; it just took (by evolutionary design) a reasonable chance on such a thing's correlating with flies.[23]

I have found myself growing more and more sympathetic to Dennett's view on this matter: he doubts that there *is* a determinate fact of the matter about the frog's mental content in such cases, and furthermore thinks this is not a serious problem.[24] Still, I think, we can take his point that more intelligence—*i.e.,* more sophisticated routes to goal satisfaction—nails down mental content *more* precisely. If the frog also had infrared sensors that needed to be triggered simultaneously with the right retinal stimulations, for example, then *dark moving spot* is no longer sufficiently explanatory for why its tongue snapped; it would have to be at least *dark, warm moving spot.* Suppose we add acute smell, acute hearing, eyes that are telescopic and high-speed (that is, with a high "critical flicker fusion" threshold), an ingrained memory bank of various sensory profiles to snatch at and not snap at, and the capacity to add to and adjust that memory bank based on experience. Each such addition means fewer plausible candidates for what the frog *thinks* it is snapping at. Dennett (1981) says

> the more we add, the richer or more demanding or specific the semantics of the system, until eventually we reach systems for which unique semantic interpretation is practically (but never in principle) dictated. ... [A]s systems become perceptually richer and behaviorally more versatile, it becomes harder and harder to make substitutions in the actual links of the system to the world without changing the organization of the system itself. If you change its environment, it will *notice*, in effect, and make a change to its internal state in response. (p. 30–31)

The suggestion here, I take it, is that mental content can be relatively constrained by multiple routes of *embedding* in the environment.[25] A frog that does not alter its behavior when its environment throws it more metal balls than flies is not particularly sensitive to the details of its environment, while one that goes seeking greener, more fly-infested pastures when bombarded with metal balls is more plausibly "thinking" about flies and "noticing" that it isn't getting any.

The examples so far have exposed a philosophical tendency to focus on indicative mental content, but I propose we take a similar lesson on the imperative side. Recall the paperclip maximizer that taped pictures to its cameras, because it

[23] Neander (2018) has a good overview of the indeterminacy problem in the context of "teleological" theories for reading mental content off of physical facts. The best example of such a theory, perhaps, is based in Ruth Garrett Millikan's seminal (1984).

[24] See *e.g.* Dennett (1987).

[25] At least, our access to and attributions of mental content will be more constrained, if not the content itself.

Note Fred Dretske (1986) takes the *learning* aspect to be especially important; as long as there are a fixed number of sensory routes $s_1, s_2 \ldots s_n$ to fly detection, we can always say what's *really* meant is "s_1, *or* s_2, *or* ... s_n" rather than "fly". But not so if the set of perceptual routes is indeterminate, depending on what the creature learns.

was rewarded when its visual stream included massive piles of paperclips. The *content* of that reward signal is unclear: is it a loose, easily subvertible directive actually to make more paperclips? Or is it a more narrow directive to gain *images* of paperclips, by hook or by crook? One way to put the question, roughly speaking, is to look at the prototypical causal chain explaining the behavior, and ask where on that chain are the content-determining causes: a distal cause, like the designers' intentions? An intermediate cause, like paperclips? Or a proximal cause, like the digitized sensory stream of paperclips?

This strikes me as a question like whether the frog is thinking about survival affordances, flies, or dark moving spots. In the frog case, I suggested that multiple low-level perceptual modes can constrain the indicative content toward the richer and appropriately intermediate cause (a *fly*). Similarly, in the RL case, perhaps the proximal content of multiple, incommensurable reward signals can triangulate on an imperative with rich and appropriately distal content. If the paperclip maximizer is rewarded not just for visual inputs of paperclips, for example, but also for the right combination with the feel of wire (or raw materials) through its intake channels, the characteristically tinkly sound the clips make as they hit the pile, and so on, then it becomes more plausible that taken together the system has a goal of *making paperclips*.

It seems to me that this is roughly the solution evolution found for us humans. On average—and despite short-circuiting opportunities—enough humans reach the distal evolutionary goal of *reproduction* through a combination of proximal rewards for eating, having sex, caring for young, and so on.[26] This is not to imply that reproduction is our one true final goal, but only the goal nature imperfectly designed us to achieve; the multiplicity of things we find rewarding together point us at least as well toward "happiness" or "life satisfaction" or some such. Of course the possibility that such goals may totally subvert nature's "intended" goal for humanity illustrates the danger here; we have to do at least as well as eons of natural selection.

What I propose, in effect, is that we provide a value learner with *multiple*, concrete, simple, and proximal final values with the aim that, through coherence reasoning, they will blend into the content of *one* abstract, complex, and distal final value. These are the agents I called *miktotelic*: "blended-goal" agents.

I think this also proposal matches our subjective experience of specifying our final values. As a kind of case study, consider the story of Howard Raiffa's difficult decision. He was an academic who at one point had to decide whether to keep his comfortable post at Columbia University, or take a new job offer from Harvard. While pacing the halls and fretting, the story goes, he ran into the philosopher of science Ernest Nagel. Nagel archly pointed out that Raiffa's academic expertise was in the relatively new field of *decision theory*. "Apply

[26] Honestly I often think of this on the simple model from a computer game I used to play (back before my own reproductive successes), *The Sims*. To keep your simulated person happy in the game requires maintaining several ever-decaying signals at once: "hunger", "social", "bladder", "hygiene", "energy" (requires enough rest), and "fun".

your own theories," Nagel in effect told Raiffa. "Crunch the numbers." To this, Raiffa infamously replied, "Come on, Ernest. This is *serious*."[27]

In point of fact, Raiffa said in an interview that he *did* apply his theories and crunch the numbers—he and his wife looked at "ten objectives which we scored and weighed."[28] (It's worth nothing, though, that after the calculations were done, they also "tested" their decision by committing in every way except formally, to see how they slept for a week.) Though few sit down to do the math, the attempt to weigh different "objectives" against each other should sound familiar. When faced with hard decisions like these, it feels as though one decision fits some of our values, another fits other of our values, and we are not sure how to trade them off. For our purposes we can imagine the Raiffas had just three objectives to trade off: perhaps support for research (including colleagues, teaching load, and interdisciplinary opportunities), material comfort (salary, benefits, and relative cost of living), and culture (including network of friends). We might imagine the scores came out something like this:

	research	comfort	culture
Harvard	7	8	4
Columbia	5	6	9

Let us call the individual objectives the "simple" values, and the complex tradeoff that the Raiffas are seeking to maximize the "complex value."[29] Such decisions are easy when one option outscores the other on *all* the simple values—but often, as here, there is no such "dominating" solution. (Raiffa explicitly says neither choice dominated.) If we simply add up the individual scores, then Columbia edges out Harvard, but Harvard wins if we count the number of simple values for which it's better. Or, like the Raiffas, we could assign weights of relative importance to the simple values, and take the weighted average: if for example they assigned weights of $\langle 5, 3, 2\rangle$ to the respective values then Harvard wins, and if they assigned weights of $\langle 3, 3, 4\rangle$ then Columbia wins. But then how are those weights to be set?

Assuming we are biological machines, there must be *some* algorithm somewhere to settle on such questions. (Anyway there would have to be one for AIs.) Of course the algorithm in question could be *arbitrary*, taking random factors of one kind or another into account, in effect flipping a mental coin. But I do not think so. Sure, *some* elements will typically be arbitrary, such as framing effects of

[27] I got this story from Thagard, who recently claims pretty good corroboration for it; see *e.g.* the opening of chapter 6 in Thagard (2010).

[28] Raiffa and Fienberg (2008) p. 142.

[29] These are not meant as actual examples of what I mean by "simple" values in humans, which I take ultimately to be biological, fixed reinforcers roughly like the "four Fs" (food, fight, flight, and reproduction). Thus a *relatively* simple value like "adventure" might itself be a complex blend of lower-level reinforcers to do with novelty and how it is registered in the brain (biologically as dopamine, or computationally as surprisal measure, *etc.*).

the question, or our mood at the time. But to say such hard choices are *entirely* arbitrary (when no option dominates) is quite a skeptical position—it suggests there can be no better or worse answers in these cases. I trust this is not our experience; we fret about playing our different objectives against each other because we think one combination will be *better* for us, and we don't know which it is. This notion that some combinations of simple values could be better or worse than others is, I suggest, what makes it the case that there really is some further, complex, underspecified value like "happiness" blended out of them.

The first challenge here is to spell out the "blending". On the one hand, the multiple simple goals must ultimately be in *some* sense reducible to one measure of overall preference, it seems, in order to result in definitive and non-arbitrary action selection.[30] On the other hand, the simple goals cannot be perfectly fungible if they are to be truly distinct. For example, if to the Raiffas more creature comfort is perfectly exchangeable for less culture and vice versa, then we may as well treat their sum as *one* disjunctively characterized value for maximizing.

Such difficulties have already been explored in the literature on *multi-objective optimization*. In multi-objective reinforcement learning, for example, the reward comes from a vector of simple reinforcers $\langle r_1, r_2, \ldots, r_n \rangle$. Like the Raiffa case above, such vectors of are generally not straightforwardly comparable, so policy selection requires some further strategy. For miktotelic purposes, the most appropriate strategy is to find a principled way to *scalarize* the vector, smashing its elements into one uber-reward number.[31] The Raiffas did this by taking a weighted average of the simple values, but there are many more complex possibilities.

As an oversimplified example, a paperclip maximizer might need a fairly consistent tactile sense of wire being fed to the twist-and-cut component, but only occasional visual inputs of piles of paperclips, and even less common sensory reassurances that there is a sufficient supply of metal in the world to continue.[32] Some constraints would also apply to relations among different component reward signals; perhaps the reward for the proprioceptive sense of having gone through a twist-and-cut motion should always outweigh visual rewards, for example. Meeting or failing these constraints might involve different kinds of rewards or penalties in the final measure; perhaps any time $r_3 < r_{17}$, the agent incurs a reward equal to 25% of r_{17}, or perhaps if there is any time interval of length n over which the total of r_6 falls below some set parameter, the agent incurs a

[30] For a nuanced discussion of such commensurability, see chapter VI of Richardson (1994).

[31] See Wang (2014) and Gábor, Kalmár, and Szepesvári (1998). Another strategy besides scalarizing is to treat each Pareto-optimal policy proposal as a kind of sub-agent with negotiating power; see Critch (2017). I might mention that yet another type of approach to reconciling multiple basic values is to elaborate the DECO model of deliberative coherence from Millgram and Thagard (1996) into a model of "belief-desire coherence"—as I have previously (2003) sought to do.

[32] This is oversimplified in part because an intelligent agent would *learn* some of these as instrumental goals.

penalty exponential in the shortfall.

So far we have considered an RL version of miktotelic agents, but similar considerations apply for miktotelic utility agents: instead of one utility function, provide a vector $\langle U_1, U_2, \ldots, U_n \rangle$ of utility functions, plus a set of constraints. In both cases, each component utility function or reward signal might be relatively simple, but determining the resulting total reward or utility *via* the constraints is computationally complex.[33] This complexity of determining the final preference ordering (to pick a term neutral between the RL and utility agent cases) is crucial—it is what makes the blended, complex value mysterious enough to require learning. *If* there is one complex phenomenon underlying all the simple imperative signals (as *fly* might underlie *dark warm buzzing ... spot*), the value learning agent will have to resort to any available information in order to approximate it.[34]

Thus suppose, in a (relatively) simple case, we wish our superintelligence to maximize human happiness. This is an abstract goal, in need of specification; Scrooge had trouble specifying it, and so do we. How could we seed it in a value-learning AI? If we just treat visual appearances of smiles as proxy evidence for happiness, then as Eliezer Yudkowsky (2011) points out, the superintelligence could "tile the future light-cone of Earth with tiny molecular smiley-faces." Clearly we would not have succeeded in a superintelligence with values that have *happiness* in their content. But if visual appearances of smiles bring defeasible reward, *and* so do audible signals of laughter, and volunteered verbal reports of happiness, and lighthearted whistling, and contented sighs, and longing gazes, and ecstatic dancing, and lack of coercion, and certain fMRI results, and so on—and if all those reward signals are set with constraints and thrown into a coherence calculation, then it may be (*may* be) that the coherently reasoning, miktotelic value learner will be forced to start theorizing about how best to balance these conflicting considerations, and at some point stumble upon the idea that there is one mysterious phenomenon underlying (enough instances of) them all, worthy of investigating.

No doubt the miktotelic approach faces its own serious challenges. The most obvious is what I think of as the *recipe problem*: it will be difficult to determine what simple values, in what arcane mixture, together blend into genuine pursuit of a complex and friendly final goal. Normally we can try to reverse engineer a complex recipe by patient trial and error. But when it comes to superintelligences, we probably won't have that luxury; our first trial (and error) is likely to be our last.

[33] I mean the reward signals or utility functions can be "simple" in the sense of low Kolmogorov complexity: essentially, they require relatively few lines of code to specify precisely. Calculating the combined total is "complex" in the different sense that, as in other weighted constraint satisfaction problems, finding the vector to optimize the scalar typically cannot be done in reasonable amounts of time (even by a superintelligence), and must be approximated.

[34] There is more to be said about when and whether there *is* an "underlying phenomenon". I will not be saying it here, though.

Even if we had complete recipes for each candidate complex friendly goal, though, we would still have to choose *which* final values we should design an agent to learn. This was our third philosophical problem for value learning, to which I now briefly turn.

Learning specific *ethical* final values

Chapter 13 of *Superintelligence* considers the question of ideal seed values in detail. As Bostrom points out, it is closely related to—but not necessarily the same thing as—asking what the *ethically correct* value system is for any agent to have.

Obviously I will not be settling the question of the *right* value system here—but I want to suggest that coherence reasoning can help, given properly seeded simple values. Though philosophers disagree on the moral facts, there is fairly broad agreement on the *method* that should ideally be used to extract them: "wide reflective equilibrium."[35] This method is basically itself a form of coherence reasoning: look at considered evaluative judgments of particular cases, and try to generalize them into principles; then, test the principles against the cases—sometimes revising the principle, and sometimes rejecting the particular judgments, depending on the overall coherence.[36]

For example, we could potentially give a miktotelic agent an array of basic reinforcements and inhibitions to correspond with our own varied and particular judgments of rightness and wrongness, and let the coherence engine determine a theory that best unifies these. It might have basic aversions to perceptions of violence, say—but then coherence calculations might determine that some particular acts of violence are justified by wider principles gleaned from other basic aversions. A superintelligence would presumably be particularly good at calculating such coherence, and perhaps come to a value system that we admire from our own perspective as clearly more coherent than our own.

In summary, then, here are the interrelated answers to the three problems with which we began.

1. An agent can *learn* a final goal by *specifying* an ambiguous, complex final goal through a coherence calculation.
2. An agent can have a *complex* final goal of fairly determinate content by building it out of simple goals *blended* with constraints on their relations.
3. An agent can learn the *right* final goal by seeding it with simple values of the type that in coherent *reflective equilibrium* will lead to plausible ethical principles.

[35] The seminal statement is in Rawls (1971), with elaboration in *e.g.* Daniels (1979).

[36] Reflective equilibrium over human value judgments seems as though it would result in something closely related to the "Coherent Extrapolated Volition" from Yudkowsky (2004); Bostrom discusses the proposal in some detail starting from loc. 4907.

Obviously, this "miktotelic" proposal for machines learning values is—like much philosophical work—just the barest outline of how to proceed. Even if it withstands criticism at the conceptual level, there is much more work to be done on the computational one.

References

Andersen, Ross. 2013. "Omens." Aeon. https://aeon.co/essays/will-humans-be-around-in-a-billion-years-or-a-trillion.

Aristotle. 350BC. *Nicomachean Ethics*. Translated by W. D. Ross. MIT Classics. http://classics.mit.edu/Aristotle/nicomachaen.html [sic].

Bostrom, Nick. 2003. "Ethical Issues in Advanced Artificial Intelligence." In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, edited by Iva Smit and George E. Lasker, 12–17. Windsor ON: International Institute for Advanced Studies in Systems Research / Cybernetics.

———. 2014. *Superintelligence: Paths, Dangers, Strategies*. Kindle edition. Oxford: Oxford University Press.

Churchland, Paul M. 1988. *Matter and Consciousness*. 1999 edition. Cambridge: MIT Press.

Critch, Andrew. 2017. "Toward Negotiable Reinforcement Learning: Shifting Priorities in Pareto Optimal Sequential Decision-Making." http://arxiv.org/abs/1701.01302.

Daniels, Norman. 1979. "Wide Reflective Equilibrium and Theory Acceptance in Ethics." *Journal of Philosophy* 76 (5): 256–82.

De Blanc, Peter. 2011. "Ontological Crises in Artificial Agents' Value Systems." San Francisco, CA: Machine Intelligence Research Institute. https://intelligence.org/files/OntologicalCrises.pdf.

Dennett, Daniel C. 1981. "True Believers: The Intentional Strategy and Why It Works." In *The Intentional Stance*, 1996 edition, 13–35. Cambridge: MIT Press.

———. 1987. "Evolution, Error, and Intentionality." In *The Intentional Stance*, 1996 edition, 287–321. Cambridge: MIT Press.

Dewey, Daniel. 2011. "Learning What to Value." San Francisco, CA: Machine Intelligence Research Institute. https://intelligence.org/files/LearningValue.pdf.

Dickens, Charles. 1843. *A Christmas Carol in Prose: Being a Ghost Story of Christmas*. London: Chapman & Hall. http://www.gutenberg.org/ebooks/46.

Dretske, Fred. 1986. "Misrepresentation." In *Belief: Form, Content, and Function*, edited by Radu J. Bogdan, 17–36. Oxford: Oxford University Press.

Everitt, Tom, and Marcus Hutter. 2016. "Avoiding Wireheading with Value Reinforcement Learning." http://arxiv.org/pdf/1605.03143v1.pdf.

Fodor, Jerry. 1990. *A Theory of Content and Other Essays*. Cambridge: MIT Press.

Friston, Karl, Francesco Rigoli, Dimitri Ognibene, Christoph Mathys, Thomas Fitzgerald, and Giovanni Pezzulo. 2015. "Active Inference and Epistemic Value." *Cognitive Neuroscience* 6 (4): 187–214.

Gábor, Zoltán, Zsolt Kalmár, and Csaba Szepesvári. 1998. "Multi-Criteria Reinforcement Learning." In *Proceedings of the Fifteenth International Conference on Machine Learning*, 197–205. ICML '98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. http://dl.acm.org/citation.cfm?id=645527.657298.

Grünwald, Peter D. 2007. *The Minimum Description Length Principle*. Cambridge: MIT Press.

Hadfield-Menell, Dylan, Stuart J. Russell, Pieter Abbeel, and Anca Dragan. 2016. "Cooperative Inverse Reinforcement Learning." In *Advances in Neural Information Processing Systems 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, 3909–17. Curran Associates, Inc. http://papers.nips.cc/paper/6420-cooperative-inverse-reinforcement-learning.pdf.

Hibbard, Bill. 2012. "Model-Based Utility Functions." *Journal of Artificial General Intelligence* 3 (1): 1–24.

Hume, David. 1739. *A Treatise of Human Nature*. 1896 edition, edited by L. A. Selby-Bigge. Oxford: Clarendon Press. https://books.google.com/books/about/A_Treatise_of_Human_Nature.html?id=5zGpC6mL-MUC.

Kolnai, Aurel. 1962. "Deliberation Is of Ends." In *Varieties of Practical Reasoning*, edited by Elijah Millgram, 259–78. MIT Press.

Lettvin, J. Y., H. R. Maturana, W. S. McCulloch, and W. H. Pitts. 1959. "What the Frog's Eye Tells the Frog's Brain." *Proceedings of the IRE* 47 (11): 1940–51. https://doi.org/10.1109/JRPROC.1959.287207.

Millgram, Elijah. 2008. "Specificationism." In *Reasoning: Studies of Human Inference and Its Foundations*, edited by Jonathan E. Adler and Lance J. Rips, 731–47. Cambridge: Cambridge University Press.

Millgram, Elijah, and Paul Thagard. 1996. "Deliberative Coherence." *Synthese* 108 (1): 63–88.

Millikan, Ruth Garrett. 1984. *Language, Thought, and Other Biological Categories*. 1995 edition. Cambridge: MIT Press.

———. 1989. "Biosemantics." In *White Queen Psychology and Other Essays for Alice*, 83–101. Cambridge: MIT Press.

Neander, Karen. 2018. "Teleological Theories of Mental Content." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2018.

https://plato.stanford.edu/archives/spr2018/entries/content-teleological/; Metaphysics Research Lab, Stanford University.

Ng, Andrew Y., and Stuart J. Russell. 2000. "Algorithms for Inverse Reinforcement Learning." In *Proceedings of the Seventeenth International Conference on Machine Learning*, 663–70. ICML '00. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. http://dl.acm.org/citation.cfm?id=645529.657801.

Nozick, Robert. 1974. *Anarchy, State, and Utopia*. Oxford: Basil Blackwell.

Petersen, Steve. 2003. "Belief-Desire Coherence." PhD thesis, University of Michigan.

———. 2017. "Superintelligence as Superethical." In *Robot Ethics 2.0*, edited by Patrick Lin, Ryan Jenkins, and Keith Abney, 322–37. New York: Oxford University Press.

Raiffa, Howard, and Stephen E. Fienberg. 2008. "The Early Statistical Years: 1947–1967 a Conversation with Howard Raiffa." *Statistical Science* 23 (1): 136–49. http://www.jstor.org/stable/27645884.

Rawls, John. 1971. *A Theory of Justice*. 1995 edition. Cambridge, MA: Harvard University Press.

Richardson, Henry S. 1994. *Practical Reasoning About Final Ends*. Cambridge: Cambridge University Press.

Sezener, Can Eren. 2015. "Inferring Human Values for Safe AGI Design." In *Artificial General Intelligence*, edited by Jordi Bieger, Ben Goertzel, and Alexey Potapov, 152–55. Switzerland: Springer International Publishing.

Shah, Rohin. 2018. "Value Learning." The AI Alignment Forum, https://www.alignmentforum.org/s/4dHMdK5TLN6xcqtyc.

Soares, Nate. 2016. "The Value Learning Problem." San Francisco, CA: Machine Intelligence Research Institute. https://intelligence.org/files/ValueLearningProblem.pdf.

Sutton, Richard S., and Andrew G. Barto. 1998. *Reinforcement Learning: An Introduction*. Cambridge: MIT Press.

Thagard, Paul. 1988. *Computational Philosophy of Science*. 1993 edition. Cambridge: MIT Press.

———. 2000. *Coherence in Thought and Action*. Cambridge: MIT Press.

———. 2010. *The Brain and the Meaning of Life*. Princeton NJ: Princeton University Press.

Thagard, Paul, and Elijah Millgram. 1995. "Inference to the Best Plan: A Coherence Theory of Decision." In *Goal-Driven Learning*, edited by Ashwin Ram and David B. Leake, 439–54. Cambridge: MIT Press.

Thagard, Paul, and Karsten Verbeurgt. 1998. "Coherence as Constraint Satisfaction." *Cognitive Science* 22 (1): 1–24.

Wang, Weija. 2014. "Multi-Objective Sequential Decision Making." PhD thesis, Université Paris Sud-Paris XI.

Yudkowsky, Eliezer. 2004. "Coherent Extrapolated Volition." San Francisco, CA: Machine Intelligence Research Institute. https://intelligence.org/files/CEV.pdf.

———. 2011. "Complex Value Systems Are Required to Realize Valuable Futures." San Francisco, CA: Machine Intelligence Research Institute. https://intelligence.org/files/ComplexValues.pdf.