The Value Learning Problem

In: Ethics for Artificial Intelligence Workshop at 25th International Joint Conference on Artificial Intelligence (IJCAI-2016) New York, NY, USA 9–15 July 2016

Nate Soares
Machine Intelligence Research Institute
nate@intelligence.org

Abstract

Autonomous AI systems' programmed goals can easily fall short of programmers' intentions. Even a machine intelligent enough to understand its designers' intentions would not necessarily *act* as intended. We discuss early ideas on how one might design smarter-than-human AI systems that can inductively learn what to value from labeled training data, and highlight questions about the construction of systems that model and act upon their operators' preferences.

1 Introduction

Standard texts in AI safety and ethics, such as Weld and Etzioni [1994] or Anderson and Anderson [2011], generally focus on autonomous systems with reasoning abilities that are complementary and not strictly superior to those of humans. Relatively little attention is given to future AI systems that may be "superintelligent" in the sense of Bostrom [2014], i.e., systems "much smarter than the best human brains in practically every field, including scientific creativity, general wisdom, and social skills." Our discussion will place a greater focus on methods and frameworks for designing robust and beneficial smarter-than-human AI systems, bracketing questions about whether such systems would have moral standing of their own.

Smarter-than-human AI systems are likely to introduce a number of new safety challenges. First, bad behavior by smarter-than-human systems can have larger and more lasting consequences; an antisocial adult is more dangerous than an antisocial child, even if the adult is as physically weak as a child. Whereas low-intelligence systems can be tested and patched over many iterations, Bostrom argues that even small errors in the first superintelligent systems could have extinction-level consequences [Bostrom, 2014]. The possible development of such systems raises the stakes for AI safety work.

Second, systems that can strictly outperform humans cognitively have less to gain from integrating into existing economies and communities. Hall [2007] has argued:

> The economic law of comparative advantage states that cooperation between individuals of differing capabilities remains mutually beneficial. [...] In other words, even if AIs become much more productive than we are, it will remain to their advantage to trade with us and to ours to trade with them.

As noted by Benson-Tilsen and Soares [forthcoming 2016], however, rational trade presupposes that agents expect more gains from trade than from coercion. Non-human species have various "comparative advantages" over humans, but humans generally exploit non-humans through force. Similar patterns can be observed in the history of human war and conquest. Whereas agents at similar capability levels have incentives to compromise, collaborate, and trade, agents with strong power advantages over others can have incentives to simply take what they want.

The upshot of this is that engineering a functioning society of powerful autonomous AI systems and humans requires that those AI systems be prosocial. The point is an abstract one, but it has important practical consequences: rational agents' interests do not align automatically, particularly when they have very different goals and capabilities.

Third, superhumanly creative and adaptive systems may arrive at what Bostrom [2014, chap. 8] calls "perverse instantiations" of their programmed goals. Wiener [1960] calls this the "Sorcerer's Apprentice" problem, after the fable of an apprentice whose enchanted broom follows instructions' letter but not their spirit.

The novelty here is not that programs can exhibit incorrect or counter-intuitive behavior, but that software agents smart enough to *understand* natural language may still *base their decisions* on misrepresentations of their programmers' intent. The idea of superintelligent agents monomaniacally pursuing "dumb"-seeming goals may sound odd, but it follows from the observation of Bostrom and Yudkowsky [2014, chap. 7] that AI capabilities and goals are logically independent.[1] Humans can fully comprehend that their "designer" (evolution) had a particular "goal" (reproduction) in mind for sex, without thereby feeling compelled to forsake contraception. Instilling one's tastes or moral values into an heir isn't impossible, but it also doesn't happen automatically.

Lastly, Bostrom and Yudkowsky [2014] point out that smarter-than-human systems may become *better* than humans

[1] Bostrom's "orthogonality thesis" can be treated as an application of Hume's [1739] observation that natural-language "is" and "ought" claims are independent.

at moral reasoning. Without a systematic understanding of how perverse instantiations differ from moral progress, how can we distinguish moral genius in highly intelligent machines from moral depravity?

Given the potential long-term impact of advanced AI systems, it would be prudent to investigate whether early research progress is possible on any of these fronts. In this paper we give a preliminary, informal survey of several research directions that we think may help address the above four concerns, beginning by arguing for *indirect* approaches to specifying human values in AI agents. We describe a promising approach to indirect value specification, *value learning*, and consider still more indirect approaches based on modeling actual and potential states of human operators.

2 Valuable Goals Cannot Be Directly Specified

We argued above that highly capable autonomous systems could have disastrous effects if their values are misspecified. Still, this leaves open the possibility that specifying correct values is *easy*, or (more plausibly) that it presents no special difficulties over and above the challenge of building a smarter-than-human AI system.

A number of researchers have voiced the intuition that some simple programmed goal would suffice for making superintelligent systems robustly beneficial. Hibbard [2001], for example, suggested training a simple learning system to recognize positive human emotions from facial expressions, voice tones, and body language. Hibbard then proposed that machines of much greater capability—perhaps even superintelligent machines—could be programmed to execute actions predicted to lead to futures with as many "positive human emotions" as possible, as evaluated by the original simple learning system.

This proposal has some intuitive appeal—wouldn't such a system always act to make humans happy?—until one considers the Sorcerer's Apprentice. We have a particular set of associations in mind when we speak of "positive human emotions," but the simple learner would almost surely have learned a different and simpler concept, such as "surface features correlating with positive human emotions in the training data." This simpler concept almost surely does not have its *maximum* at a point which Hibbard would consider to contain lots of positive human emotions. The maximum is much more likely to occur in (for example) scenarios that contain an enormous number of tiny human-shaped animatronics acting out positive human emotions. Thus, a powerful learning system that takes actions according to how well the simple learner would rank them is liable to spend time and resources creating animatronics rather than spending time and resources making humans happy. Indeed, Hibbard [2012] himself comes to the conclusion that his proposal fails to exclude the possibility that lifelike animatronic replicas of happy people could be counted as exhibiting "positive emotions."

As another example, Schmidhuber [2007] proposes that creativity, curiosity, and a desire for discovery and beauty can be instilled by creating systems that maximize a different simple measure: "create action sequences that extend the observation history and yield previously unknown / unpredictable but quickly learnable algorithmic regularity or compressibility."

However, while it is quite plausible that human creativity and discovery are related to the act of compressing observation, an agent following Schmidhuber's goal would not behave in intuitively curious and creative ways. One simple way to meet Schmidhuber's desideratum, for example, is to appropriate resources and construct artifacts that generate cryptographic secrets, then present the agent with a long and complex series of observations encoded from highly regular data, and then reveal the secret to the agent, thereby allowing the agent to gain enormous compression on its past sensory data. An agent following Schmidhuber's goal is much more likely to build artifacts of this form than it is to pursue anything resembling human creativity. The system may not take this action in particular, but it will take actions that generate *at least that much* compression of its sensory data, and as a result, the system is unlikely to be prosocial.

Building an agent to do something which (in humans) correlates with the desired behavior does not necessarily result in a system that acts like a human. The general lesson we draw from cases like these is that most goals that are simple to specify will not capture all the contextual complexities of real-world human values and objectives [Yudkowsky, 2011]. Moral psychologists and moral philosophers aren't locked in decades- and centuries-long debates about the right codifications of ethics because they're missing the obvious. Rather, such debates persist for the simple reason that morality is complicated. People want lots of things, in very particular ways, and their desires are context-sensitive.

Imagine a simplified state space of possibilities that vary in *count* (how many happy human-shaped objects exist), in the *size* of the average happy human-shaped object, and in the average *moral worth* of happy human-shaped objects. Human experience has occurred in a small region of this space, where almost all human-shaped objects emitting what looks like happiness are $\approx$ 2-meter-sized humans with moral weight. But the highest scores on the count axis occur in tandem with low size, and the smallest possible systems that can mimic outward signs of emotion are of low moral worth.

In linear programming, it is a theorem that the maximum of an objective function occurs on a vertex of the space. (Sometimes the maximum will be on an edge, including its vertices.) For intuitively similar reasons, the optimal solution to a goal tends to occur on a vertex (or edge, or hyperface) of the possibility space. Hibbard's goal does not contain any information about size or moral worth, and so agents pursuing this goal only consider size and moral worth insofar as they pertain to pushing toward the hyperface of maximum count. To quote Russell [2014]:

> A system that is optimizing a function of n variables, where the objective depends on a subset of size $k < n$, will often set the remaining unconstrained variables to extreme values; if one of those unconstrained variables is actually something we care about, the solution found may be highly undesirable.

The Sorcerer's Apprentice problem arises when systems' programmed goals do not contain information about all relevant

dimensions along which observations can vary. The agent has been directed towards the wrong hyperface of the possibility space.[2]

When confronted with this type of failure, many have an impulse to patch the flawed goals. If Hibbard's system would make smiling animatronics, then find ways to require that the emotions come from actual humans; if the system would then put humans in a drugged stupor in order to make them smile, forbid it from using drugs; and so on.

Such constraints cut off particular means by which the system can get a higher count, but they don't address the underlying problem that the system is still maximizing count. If one causal pathway is forbidden, then the system will follow the nearest non-forbidden causal path—e.g., mechanically manipulating the pleasure centers of human brains. It isn't feasible to patch every goal; nor is it safe to patch as many as come to mind and assume that there are no unforeseen perverse instantiations.

Intuitively, we would like to direct the intelligence of highly advanced systems to solving some of this problem on our behalf, and we would like such systems to attend to our likely intentions even when our formal and informal representations of our intentions are flawed. The notion of the operator's "intentions," however, is unlikely to lend itself to clean formal specification. By what methods, then, could an intelligent machine be constructed to reliably learn what to value and to act as its operators intended?

3 Inductive Value Learning

Correctly specifying a formal criterion for recognizing a cat in a video stream by hand is difficult, if not impossible. This does not mean, however, that cat recognition is hopeless; it means that a level of indirection is required. An image recognition system can be constructed and trained to recognize cats. We propose that the value learning problem be approached by similarly indirect means.

Inductive value learning via labeled training data raises a number of difficulties. A visual recognition system classifies images; an inductive value learning system classifies *outcomes*. What are outcomes? What format would a value-learning data set come in?

Imagine a highly intelligent system that uses large amounts of data to construct a causal model of its universe. Imagine also that this world-model can be used to reason about the likely outcomes of the agent's available actions, that the system has some method for rating outcomes, and that it executes the action leading to the most highly rated outcome. In order for the system to inductively learn what to value, the system must be designed so that when certain "training" observations are made (or specially-demarcated updates to its world-model occur), labeled training data extracted from the observation or update alters the method by which the system ranks various potential outcomes.

[2]Instead of trying to direct the system toward exactly the right hyperface, one might try to create a "limited optimization" system that doesn't push so hard in whatever direction it moves. This seems like a promising research avenue, but is beyond the scope of this paper.

This simple model highlights a central concern and two open questions relevant to inductive value learning.

3.1 Corrigibility

Imagine that some of an agent's available actions allow it to modify itself, and that it currently assigns high utility to outcomes containing high numbers of animatronic replicas of humans. It may be the case that, according to the system's world-model, all of the following hold: (1) if more training data is received, those high-rated outcomes will have their ratings adjusted downwards; (2) after the ratings are adjusted, the system will achieve outcomes that have fewer cheap animatronics; and (3) there are actions available which remove the inductive value learning framework.

In this situation, a sufficiently capable system would favor actions that disable its value learning framework. It would not necessarily consider its own process of learning our values a good thing, any more than humans must approve of psychological disorders they possess. One could try to construct protected sections of code to prevent the value learning framework from being modified, but these constraints would be difficult to trust if the system is more clever than its designers when it comes to exploiting loopholes.

A robustly safe initial system would need to be constructed in such a way that actions which remove the value learning framework are poorly rated even if they are available. Some preliminary efforts toward describing a system with this property have been discussed under the name *corrigibility* by Soares and Fallenstein [2015], but no complete proposals currently exist.

3.2 Ontology Identification

The representations used in a highly intelligent agent's world-model may change over time. A fully trustworthy value learning system would need to not only classify potential outcomes according to their value, but persist in doing so correctly even when its understanding of the space of outcomes undergoes a major change.

Consider a programmer that wants to train a system to pursue a very simple goal: produce diamond. The programmers have an atomic model of physics, and they generate training data labeled according to the number of carbon atoms covalently bound to four other carbon atoms in that training outcome. For this training data to be used, the classification algorithm needs to identify the atoms in a potential outcome considered by the system. In this toy example, we can assume that the programmers look at the structure of the initial world-model and hard-code a tool for identifying the atoms within. What happens, then, if the system develops a nuclear model of physics, in which the ontology of the universe now contains primitive protons, neutrons, and electrons instead of primitive atoms? The system might fail to identify any carbon atoms in the new world-model, making the system indifferent between all outcomes in the dominant hypothesis. Its actions would then be dominated by any tiny remaining probabilities that it is in a universe where fundamental carbon atoms are hiding somewhere.

This is clearly undesirable. Ideally, a scientific learner should be able to infer that nuclei containing six protons are the

true carbon atoms, much as humans have done. The difficulty lies in formalizing this process.

To design a system that classifies potential outcomes according to how much diamond is in them, some mechanism is needed for identifying the intended ontology of the training data within the potential outcomes as currently modeled by the AI. This is the *ontology identification* problem introduced by de Blanc [2011] and further discussed by Soares [2015].

This problem is not a traditional focus of machine learning work. When our only concern is that systems form better world-models, then an argument can be made that the nuts and bolts are less important. As long as the system's new world-model better predicts the data than its old world-model, the question of whether diamonds or atoms are "really represented" in either model isn't obviously significant. When the system needs to consistently pursue certain outcomes, however, it matters that the system's internal dynamics preserve (or improve) its representation of which outcomes are desirable, independent of how helpful its representations are for prediction. The problem of making correct choices is not reducible to the problem of making accurate predictions.

Inductive value learning requires the construction of an outcome-classifier from value-labeled training data, but it also requires some method for identifying, inside the states or potential states described in its world-model, the referents of the labels in the training data.

This could perhaps be done during the course of inductive value learning. The system's methods for inferring a causal world-model from sense data could perhaps be repurposed to infer a description of what has been labeled. If the system adopts a better world-model, it could then re-interpret its training data to re-bind the value labels.

This looks like a promising line of research, but it seems to us to require new insights before it is close to being formalizable, let alone usable in practice. In particular, we suspect that ontology identification will require a better understanding of algorithms that construct multi-level world-models from sense data.

3.3 Ambiguity Identification

Reinforcement learning can be thought of as a method for sidestepping these difficulties with value learning. Rather than designing systems to learn which outcomes are desirable, one creates a proxy for desirable outcomes: a reward function specified in terms of observations. By controlling rewards via a reward signal, the operator can then judiciously guide the learner toward desired behaviors. Indirect proxies for desired outcomes, however, face many of the same Sorcerer's Apprentice difficulties. Maximizing how often an operator transmits a reward signal is distinct from the problem of maximizing the operator's satisfaction with outcomes; these goals may coincide in testing environments and yet diverge in new environments—e.g., once the learner has an opportunity to manipulate and deceive its operator or otherwise hijack its reward channel [Bostrom, 2014, chap. 12]. For further discussion, see Soares [2015].

Superintelligent systems that achieve valuable real-world outcomes may need goals specified in terms of desirable outcomes, rather than rewards specified in terms of observations. If so, then we will need some robust way of ensuring that the system learns our goals, as opposed to superficially similar goals.

When training a recognition system, producing satisfactory training data is often a difficult task. There is a classic parable of machine learning (told by, e.g., Dreyfus and Dreyfus [1992]) of an algorithm intended to classify whether or not pictures of woods contained a tank concealed between the trees. Pictures of empty woods were taken one day; pictures with concealed tanks were taken the next. The classifier identified the latter set with great accuracy, and tested extremely well on the portion of the data that had been withheld from training. However, the system performed poorly on new images. It turned out that the first set of pictures had been taken on a sunny day, while the second set had been taken on a cloudy day. The classifier was not identifying tanks; it was identifying image brightness!

The same mistake is possible when constructing a training data set for inductive value learning. In value learning, however, such mistakes may be more difficult to notice and more consequential. Consider a training set that successfully represents real-world cases of happy human beings (labeled with high ratings) and real-world cases of pointless human suffering (rated poorly). The simplest generalization from this data may, again, be that human-shaped-things-proclaiming-happiness are of great value, even if these are animatronics imitating happiness. It seems plausible that someone training an inductive value learner could neglect to include a sufficiently wide variety of animatronics mimicking happiness and labeled as low-value. How many other obvious-in-retrospect pitfalls are hiding in our blind spots?

A training set covering all relevant dimensions that we can think of may yet exclude relevant dimensions. A robustly safe value learner would need to be able to identify *new plausibly-relevant dimensions along which no training data is provided*, and query the operators about these ambiguities. This is the kind of modification that would help in actually solving the value learning problem, as opposed to working around it. At the same time, this is the kind of modification that could take advantage of machines' increased capabilities as the field of AI advances.

Formalizing this idea is a key open problem. Given a data set which classifies outcomes in terms of some world-model, how can dimensions along which the data set gives little information be identified? One way to approach the problem is to study how humans learn concepts from sparse data, as discussed by Tenenbaum *et al.* [2011] and Sotala [2015]. Alternatively, it may be possible to find some compact criterion for identifying ambiguities in a simpler fashion. In both cases, further research could prove fruitful.

4 Modeling Intent

The problem of ambiguity identification may call for methods beyond the inductive learning of value from training data. An intelligent system with a sufficiently refined model of humans may already have the data needed, provided that the right question is asked, to deduce that humans are more likely to care about whether happy-looking human-shaped things have brains than about the nearby breeze. The trouble would be

designing the system to use this information in exactly the right way.

Picture a system that builds multi-level environment models from sensory data and learns its values inductively. One could then specially demarcate some part of the model as the "model of the operator," define some explicit rules for extracting a model of the operator's preferences from the model of the operator (in terms of possible outcomes), and adjust the ratings on various outcomes in accordance with the model of the operator's preferences. This would be a system which attempts to learn and follow another agent's intentions, as opposed to learning from labeled training data—a "do what I mean" (DWIM) architecture.

The inverse reinforcement learning (IRL) techniques of Ng and Russell [2000] can be viewed as a DWIM approach, in which an agent attempts to identify and maximize the reward function of some other agent in the environment. However, existing IRL formalizations do not capture the full problem; the preferences of humans cannot necessarily be captured in terms of observations alone. For example, a system, upon observing its operator lose at a game of chess, should not conclude that its operator wanted to lose at chess, even if the system can clearly see where the operator "decided" to make a bad move instead of a good one. Or imagine a human operator who has a friend that must be put into hiding. The learner may either take the friend to safety, or abandon the friend in a dangerous location and use the resources saved in this way to improve the operator's life. If the system reports that the friend is safe in both cases, and the human operator trusts the system, then the latter observation history may be preferred by the operator. However, the latter outcome would definitely not be preferred by most people if they had complete knowledge of the outcomes.

Human preferences are complex, multi-faceted, and often contradictory. Safely extracting preferences from a model of a human would be no easy task. Problems of ontology identification recur here: the framework for extracting preferences and affecting outcome ratings needs to be robust to drastic changes in the learner's model of the operator. The special-case identification of the "operator model" must survive as the system goes from modeling the operator as a simple reward function to modeling the operator as a fuzzy, ever-changing part of reality built out of biological cells—which are made of atoms, which arise from quantum fields.

DWIM architectures must avoid a number of other hazards. Suppose the system learns that its operator model affects its outcome ratings, and the system has available to it actions that affect the operator. Actions which manipulate the operator to make their preferences easier to fulfill may then be highly rated, as they lead to highly-rated outcomes (where the system achieves the operator's now-easy goals). Solving this problem is not so simple as forbidding the system from affecting the operator; any query made by the system to the operator in order to resolve some ambiguity will affect the operator in some way.

A DWIM architecture requires significant additional complexity on top of inductive value learning: the agent's goal-adjusting learning system no longer simply classifies outcomes; it must also model humans and extract human preferences about human-modeled outcomes, and translate between human-modeled future outcomes and future outcomes as modeled by the system. The hope is that this complexity purchases a system that potentially achieves full and direct coverage of the complexity of human value, without relying on the abilities of the programmers to hand-code exceptions for every edge case or compose exactly the right training set. Further investigations into inverse reinforcement learning or other methods of constructing satisfactory initial operator models may be a good place to start studying the plausibility of DWIM architectures.

5 Extrapolating Volition

A DWIM architecture may be sufficient when constructing a system that reliably pursues "concrete" goals (such as "cure cancer and then await instruction"), but it may not be sufficient for more complex or sophisticated goals where the operators themselves do not know what they intend—for example, "Do what I *would* want, if I had more knowledge and more time to think." None of the frameworks discussed so far seem powerful enough to specify philosophical ideas like the "ideal advisor theory" of Rosati [1995] or the "reflective equilibrium" of Rawls [1971]. Here, even "indirect" approaches to making robust and beneficial AI systems run aground of actively debated questions in moral philosophy.

One possible approach to resolving normative uncertainty (e.g., about what the operators would want if they were wiser or better people) would be to build a DWIM system that takes a model of a human operator and *extrapolates* it in the direction of e.g. Rawls' reflective equilibrium. For example, the extrapolation might predict what the operator would decide if they knew everything the system knows, or if they had considered many possible moral arguments [Bostrom, 2014, chap. 13].

However, a high-powered system searching for moral arguments that would put the operators into a reflectively stable state (as a computational expedient to fully simulating the operators' process of reflection) introduces a new set of potential pitfalls. A high-powered search for the *most* persuasive moral arguments that elicit retrospective approval of moral changes might find arguments that induce psychotic breakdowns or religious conversions. The system should be constrained to search for only "valid" moral arguments, but defining what counts as a valid moral argument is itself a major area of normative uncertainty and disagreement.

In this domain, querying for ambiguities is difficult. In everyday practice, an argument that is persuasive to smart and skeptical humans is often valid, but a superintelligent search for persuasive arguments may well discover invalid but extremely persuasive arguments.

It is difficult to identify technical approaches to indirect normativity that are tractable today, although there have been a few initial forays. Christiano [2014] informally proposes one mechanism by which a system could perhaps safely extrapolate the volition of its operator. MacAskill [2014] has given an extensive report on "meta-normativity," touching upon many different philosophical aspects of the difficulties of resolving normative uncertainty. This is an area where further philo-

sophical study may make it clearer how to begin approaching the associated long-run engineering problems.

6 Discussion

Just as human intelligence has allowed us to develop tools and strategies by which we can control our environment, so too could superintelligent systems develop tools and strategies more powerful than our own, and gain correspondingly greater control over future outcomes [Bostrom, 2014, chap. 6]. Although it is not clear how long the development of smarter-than-human systems will take, or what approaches in AI or other disciplines may prove most relevant to developing such systems, early efforts in this area are justified by its importance and neglectedness.

In the introduction to this paper, we discussed four different ways in which the potential development of superintelligent machines changes the task of AI safety and ethics work. Addressing all these concerns does not seem easy. Designs for AI systems that are intended to become superintelligent will need to be corrigible in the sense of Soares and Fallenstein [2015], i.e., willing to assist their operators in attempted corrections. The systems will need some method for learning and adopting prosocial preferences, in light of the fact that we cannot expect arbitrary rational actors to exhibit prosocial behavior in the face of large power disparities. Operators will require methods for robustly communicating their intentions to the system, if Sorcerer's Apprentice scenarios are to be avoided. And eventually, explicit methodologies for resolving normative uncertainty may be required.

This paper has given a cursory overview of a number of potential lines of research for AI value specification. We discuss these ideas in part to give an overview of plausible approaches to the concerns outlined above, and also because these are topics that seem amenable to research starting sooner rather than later, even in the face of great uncertainty about the particular architectures of future AI systems. It is difficult to know which lines of research will pan out, and we hope that this survey inspires research along a number of new paths, so that we have a firm theoretical grasp of how systems could reliably and safely learn our values in principle before it comes time to build systems that must do so in practice.

References

[Anderson and Anderson, 2011] Michael Anderson and Susan Leigh Anderson, editors. *Machine Ethics*. Cambridge University Press, 2011.

[Benson-Tilsen and Soares, forthcoming 2016] Tsvi Benson-Tilsen and Nate Soares. Formalizing convergent instrumental goals. 2nd International Workshop on AI, Ethics and Society at AAAI-2016, forthcoming 2016.

[Bostrom and Yudkowsky, 2014] Nick Bostrom and Eliezer Yudkowsky. The ethics of artificial intelligence. In Keith Frankish and William M. Ramsey, editors, *The Cambridge Handbook of Artificial Intelligence*. Cambridge University Press, New York, 2014.

[Bostrom, 2014] Nick Bostrom. *Superintelligence*. Oxford University Press, 2014.

[Christiano, 2014] Paul Christiano. Specifying "enlightened judgment" precisely (reprise). *Ordinary Ideas*, 2014.

[de Blanc, 2011] Peter de Blanc. Ontological crises in artificial agents' value systems. Technical Report arXiv:1105.3821 [cs.AI], The Singularity Institute, San Francisco, CA, 2011.

[Dreyfus and Dreyfus, 1992] Hubert L. Dreyfus and Stuart E. Dreyfus. What artificial experts can and cannot do. *AI & Society*, 6(1):18–26, 1992.

[Hall, 2007] John Storrs Hall. *Beyond AI*. Prometheus Books, 2007.

[Hibbard, 2001] Bill Hibbard. Super-intelligent machines. *ACM SIGGRAPH Computer Graphics*, 35(1):13–15, 2001.

[Hibbard, 2012] Bill Hibbard. The error in my 2001 VisFiles column, 2012.

[Hume, 1739] David Hume. *A Treatise of Human Nature*. Printed for John Noon, at the White-Hart, near Mercer's-Chapel, in Cheapside., 1739.

[MacAskill, 2014] William MacAskill. *Normative Uncertainty*. PhD thesis, St Anne's College, University of Oxford, 2014.

[Ng and Russell, 2000] Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In Pat Langley, editor, *17th International Conference on Machine Learning (ICML-'00)*, pages 663–670. Morgan Kaufmann, 2000.

[Rawls, 1971] John Rawls. *A Theory of Justice*. Belknap, 1971.

[Rosati, 1995] Connie S. Rosati. Persons, perspectives, and full information accounts of the good. *Ethics*, 105(2):296–325, 1995.

[Russell, 2014] Stuart J. Russell. Of myths and moonshine. *Edge*, 2014.

[Schmidhuber, 2007] Jürgen Schmidhuber. Simple algorithmic principles of discovery, subjective beauty, selective attention, curiosity and creativity. In Vincent Corruble, Masayuki Takeda, and Einoshin Suzuki, editors, *Discovery Science*, number 4755 in Lecture Notes in Computer Science, pages 26–38. Springer, 2007.

[Soares and Fallenstein, 2015] Nate Soares and Benja Fallenstein. Questions of reasoning under logical uncertainty. Technical Report 2015–1, Machine Intelligence Research Institute, 2015.

[Soares, 2015] Nate Soares. Formalizing two problems of realistic world-models. Technical Report 2015–3, Machine Intelligence Research Institute, 2015.

[Sotala, 2015] Kaj Sotala. Concept learning for safe autonomous AI. 1st International Workshop on AI and Ethics at AAAI-2015, 2015.

[Tenenbaum *et al.*, 2011] Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011.

[Weld and Etzioni, 1994] Daniel Weld and Oren Etzioni. The first law of robotics (a call to arms). In Barbara Hayes-Roth and Richard E. Korf, editors, *12th National Conference on Artificial Intelligence (AAAI-1994)*, pages 1042–1047, Menlo Park, CA, 1994. AAAI Press.

[Wiener, 1960] Norbert Wiener. Some moral and technical consequences of automation. *Science*, 131(3410):1355–1358, 1960.

[Yudkowsky, 2011] Eliezer Yudkowsky. Complex value systems in Friendly AI. In *Artificial General Intelligence. 4th International Conference, AGI 2011*, pages 388–393. Springer, 2011.