# The Value Learning Problem

**Nate Soares**
Machine Intelligence Research Institute
nate@intelligence.org

## Abstract

A superintelligent machine would not automatically act as intended: it will act as programmed, but the fit between human intentions and formal specification could be poor. We discuss methods by which a system could be constructed to learn what to value. We highlight open problems specific to inductive value learning (from labeled training data), and raise a number of questions about the construction of systems which model the preferences of their operators and act accordingly.

## 1 Introduction

Consider a superintelligent system, in the sense of Bostrom (2014), tasked with curing cancer by discovering some process which eliminates cancerous cells from a human body without causing harm to the human (no easy task to specify in its own right). The resulting behavior may be quite unsatisfactory. Among the behaviors not ruled out by this goal specification are stealing resources, proliferating robotic laboratories at the expense of the biosphere, and kidnapping human test subjects.

The intended goal, hopefully, was to cure cancer without doing any of those things, but computer systems do not automatically act as intended. Even a system smart enough to figure out what was intended is not compelled to act accordingly: human beings, upon learning that natural selection "intended" sex to be pleasurable only for purposes of reproduction, do not thereby conclude that contraceptives are abhorrent. While one should not anthropomorphize natural selection, humans are capable of understanding the process which created them while being unmotivated to alter their preferences accordingly. For similar reasons, when constructing an artificially intelligent system, it is not sufficient to construct a system intelligent enough to understand human intentions; the system must also be purposefully constructed to pursue them (Bostrom 2014, chap. 8).

How can this be done? Human goals are complex, culturally laden, and context-dependent. Furthermore, the notion of "intention" itself may not lend itself to clean formal specification. By what methods could an intelligent machine be constructed to reliably learn what to value and to act as its operators intended?

A superintelligent machine would be useful for its ability to find plans that its programmers never imagined, to identify shortcuts that they never noticed or considered. That capability is a double-edged sword: a machine that is extraordinarily effective at achieving its goals might have unexpected negative side effects, as in the case of robotic laboratories damaging the biosphere. There is no simple fix: a superintelligent system would need to learn detailed information about what is and isn't considered valuable, and be motivated by this knowledge, in order to safely solve even simple tasks.

This *value learning problem* is the focus of this paper. Section 2 discusses an apparent gap between most intuitively desirable human goals and attempted simple formal specifications. Section 3 explores the idea of frameworks through which a system could be constructed to learn concrete goals via induction on labeled data, and details possible pitfalls and early open problems. Section 4 explores methods by which systems could be built to safely assist in this process.

Given a system which is attempting to act as intended, philosophical questions arise: How could a system learn to act as intended when the operators themselves have poor introspective access to their own goals and evaluation criteria? These philosophical questions are discussed briefly in Section 5.

A superintelligent system under the control of a small group of operators would present a moral hazard of extraordinary proportions. Is it possible to construct a system which would act in the interests of not only its operators, but of all humanity, and possibly all sapient life? This is a crucial question of philosophy and ethics, touched upon only briefly in Section 6, which also motivates a need for caution and then concludes.

Research supported by the Machine Intelligence Research Institute (intelligence.org). Published as Technical report 2015–4.

## 2 Valuable Goals Cannot Be Directly Specified

Many people have an intuition that superintelligent systems with simple goals can lead to desirable outcomes. For example, consider Schmidhuber (2007), who suggests that creativity, curiosity, and a desire for discovery and beauty can be instilled by creating systems that maximize the degree to which past sensory data can be compressed. “The agent’s goal should be: create action sequences that extend the observation history and yield previously unknown / unpredictable but quickly learnable algorithmic regularity or compressibility” (Schmidhuber 2007).

However, most goals that are simple to specify will not capture all the complexities of human endeavors. While it is true that human creativity and discovery are related to the act of compressing observation, an agent following Schmidhuber’s goal would not be the curious and creative artificial citizen that may spring to mind. For example, one simple way to “extend the observation history and yield previously unknown / unpredictable but quickly learnable algorithmic regularity” 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. An agent following Schmidhuber’s goal is much more likely to build artifacts of this form than it is to pursue anything akin to human creativity.

Building an agent to do something which, in humans, correlates with the desired behavior, is not likely to result in a system that acts in a human manner. Instead, it is likely to result in an agent with very strange incentives.

Consider Hibbard (2001), who suggested training a simple system to recognize positive human emotions from facial expressions, voice tones, and body language. Then an intelligent system could be constructed to take actions which are predicted to lead to many positive human emotions (as recognized by the recognizer). This may seem intuitively desirable: wouldn’t such a system always act to make humans happy? Unfortunately, a system with Hibbard’s goals would not exhibit the intended behavior: actions that lead to many positive human emotions (as recognized by the recognizer) would mostly entail the production of many, many cheap animatronics mimicking positive human emotions in order to trigger the simple recognizer as much as possible.[1]

Complex goals are required to specify complex values (Yudkowsky 2011). Imagine a simplified state space of possibilities that a system could achieve, with three axes: (1) **count** of human-shaped objects emitting what looks like positive emotion; (2) **size** of average human-shaped object emitting what looks like positive emotion; and (3) average **moral worth** of human-shaped objects emitting what looks like positive emotion. Most of human experience has occurred in a small region of this space, the region where almost all human-shaped objects emitting what looks like positive emotion 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 vertexes.) 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.” Bostrom (2014, chap. 8) calls this phenomenon “perverse instantiation.” The problem is not that the agent is deliberately misinterpreting its goals, the problem is that the goals do not contain information about all relevant dimensions; the agent has been directed towards the wrong hyperface of the possibility space.[2]

When confronted with the possibility of perverse instantiation, many have an impulse to patch the flawed goals. If Hibbard’s system would make animatronics, then require that the emotions come from actual humans. If the system would then drug humans, then forbid it from using drugs. Such constraints cut off causal pathways that the system could use to get a higher count, but they don’t address the problem that the system is still maximizing count. If one causal pathway is forbidden, then the system will follow the nearest non-forbidden neighboring causal path (such as directly manipulating the pleasure centers of human brains). It is neither feasible nor safe to attempt to patch the goals until it seems that all causal pathways to imagined bad outcomes have been ruled out. The system would still be searching for ways to achieve outcomes that humans regard as bad. Even if the search is expected to come up empty, constructing such a system is imprudent.

Simple genetic algorithms can already achieve solutions in unexpected ways. Consider the algorithm of Bird and Layzell (2002), tasked with the development of an oscillating circuit. The algorithm would have produced a standalone oscillating circuit in test

[1] Hibbard has since acknowledged this flaw in his suggestion (Hibbard 2012).

[2] Instead of trying to direct the system toward exactly the right hyperface, why not 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.

scenarios, when run on an abstract representation of a circuit board containing all the features of the possibility space that *humans* thought were relevant. But when the algorithm operated in reality, it made use of dimensions of the possibility space that humans never considered: It repurposed the printed circuit tracks on its motherboard as a makeshift radio, which it used to amplify signals from nearby machines.

Even if the optimal way to satisfy the goals among all possibilities that *humans* have imagined seems like a high-value state, we cannot guarantee that we have visualized all possible candidates, and the actual solution found by superintelligent search may not be high-value. The highest scoring solution relative to Schmidhuber's compression-of-sensory-information goal corresponds not to the "typical" case (pursuing discovery) but to a "weird" edge-case (constructing artifacts that produce encryptions of regular data and then reveal the encryption key). This is the problem of *unforeseen maximums*: a fit between a formal goal and a high-value solution is difficult to achieve and difficult to verify.

The *fragility of value thesis* (Yudkowsky 2011) states that human values are both *complex* (in the sense that there is not a simple seed such as "compress sensory information" from which all components of value follow) and *fragile* (in the sense that a failure along one dimension, such as "too little respect for freedom," could destroy nearly all the available value). Motivating this thesis is beyond the scope of this paper, but assuming this thesis is correct, there are many dimensions of value (and relations between them) that must be precisely specified in order to formalize a goal which (when pursued) leads to a valuable outcome. Some dimensions may be difficult to identify, and each dimension might present a separate challenge of perverse instantiation. Assuming that value is fragile, correctly specifying what to value by hand is difficult if not impossible.

## 3 Inductive Value Learning

Correctly specifying a formal criterion for recognizing a cat in a video stream by hand is difficult if not impossible, but that does not mean that cat recognition is hopeless. It means that a level of indirection is required: a trainable recognition system can be constructed and trained to recognize cats. The value learning problem could be approached using a similar sort of indirection.

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?[3]

Imagine a powerfully intelligent system which takes significant amounts of data and builds a causal model of its universe. Imagine also that this world model can be used to reason about the probable outcomes achievable via the agent's available actions, and that the system has some method for rating outcomes and is constructed to execute the action leading to the best outcome. In order for the system to inductively learn what to value, the system must be designed so that, when certain observations are made (or certain updates to the world-model happen), labeled training data extracted from the observation or update modifies the ratings assigned to various potential outcomes.

This is a simplification, to be sure, but it highlights a central concern and two open questions relevant to inductive value learning.

### 3.1 Corrigibility

Imagine that some of the available actions allow the system to modify itself, and that it currently assigns high utility to outcomes which contain many animatronic faux-humans mimicking happiness. 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.

This system might execute actions which remove its value learning framework. It would not automatically consider its own value learning framework a good thing; if it were constructed to execute the highest-rated actions then it would simply execute the highest-rated actions. 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 once the system is sufficiently powerful to model itself and consider self-modification. In the name of safety, the initial system should 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 heading of *corrigibility* by Soares and Fallenstein (2015), but no complete proposals currently exist.

### 3.2 Ontology Identification

The representations used in the system's world model may shift over time. The inductive value learning must result in a system which not only classifies potential outcomes according to their value, but which continues

[3] Why not specify a reward function in terms of observations, allowing the user to control rewards via a reward signal and reinforcement learning techniques? Given sufficient intelligence, this would result in a system which behaves as intended only until it can gain decisive control over its reward channel (Bostrom 2014, chap. 12). In order to construct a superintelligent system that achieves valuable real-world outcomes, the system must have goals specified in terms of desirable outcomes rather than rewards specified in terms of observation. (For further discussion, see Soares [2015].)

to do so correctly even when the structure of a potential outcome undergoes a drastic shift.

An example helps to make the difficulty clear. Imagine programmers training 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. In order for this training data to be used, the classification algorithm needs to identify the atoms in a potential outcome considered by the system—say that the programmers look at the structure of the initial world model and hard-code a tool for identifying the atoms within. Now what happens when the system develops a nuclear model of physics, in which the ontology of the universe now contains no primitive atoms but instead primitive protons, neutrons, and electrons? 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 the tiny remaining probabilities that it is in a universe where fundamental carbon atoms are hiding somewhere.

This is clearly undesirable: ideally, a system should be able to deduce that nuclei containing six protons are the true carbon atoms, but how can this be done? In order to create a system which classifies potential outcomes according to how much diamond is in them, the system needs some mechanism 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).

Inductive value learning requires more than the construction of an outcome-classifier from value-labeled training data. The system must also have some method for identifying, inside the states or potential states described in its world-model, the referents of the labels in the training data. Furthermore, this method needs to be robust against changes in the world-model (de Blanc 2011).

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. Then, if the system infers a better world-model underlying all sense data, it could re-interpret the training data to re-bind the value labels.

This seems like a promising research approach, but it seems to us to require new ideas before it is close to being formalizable, let alone usable in practice. In particular, we suspect that ontology identification is likely to require a better understanding of algorithms which build multi-level world models from sense data—we usually don't think about most things in our environment on the atomic level, and these higher levels of representation seem relevant for describing goals. With a better understanding of multi-level representations, one could study methods for reliably identifying the intended referents of a label at the right level in a world model (even if the states labeled in the training data do not cleanly correspond to any particular level in the world model of the system). At present, any research into the construction of multi-level world models from sense data could yield progress.

### 3.3 Unforeseen Inductions

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. The first set of pictures were taken on a sunny day, whereas the latter were 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. With value learning, however, the mistakes could be both dangerous and difficult to notice.

Consider a training set which successfully references 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, and this may lead the system to construct animatronics imitating happiness. It seems entirely plausible that someone attempting inductive value learning could neglect to put in any observations of animatronic things mimicking happiness and labeled as low-value. How many other "too obvious" insights are hiding squarely in our anthropocentric blind spots?

Concerns about perverse instantiation arise again when constructing a training set for inductive value learning: a training set that covers all relevant dimensions that *we* can think of may not cover all relevant dimensions. If an inductive value learner is to be safe, the system needs 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 another 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?

Ambiguity identification may be difficult to do correctly. It is easy to imagine a system which receives value-labeled training data, and then spends the first week querying about wind patterns, and spends the second week querying about elevation differentials, only to query whether brains are necessary long after the programmers lost interest. There is an intuitive sense in which humans "obviously" care about whether the human-shaped-things have brains more than we care about whether the people are on mountains, but it may not be obvious to the system.

One way to approach the problem is to study how humans learn concepts from sparse data, as discussed by Tenenbaum et al. (2011) and further by Sotala (forthcoming). Alternatively, it may be possible to find some other compact criterion for identifying ambiguities in a simpler fashion. In both cases, further research into ambiguity identification could prove fruitful.

## 4 Acting as Intended

The problem of ambiguity identification may motivate the need for methods beyond inductive value learning. An intelligent system with a sufficiently refined model of humans may already have the data needed, given 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 breezes nearby. The trouble would be designing the system to use this information in exactly the right way.

As before, picture a system that builds multi-level environment models from sense data and has a framework for inductive value learning. One could then specially demarcate some part of the model as the "model of the operator," define some explicit rules for extracting the model of the preferences from the model of the operator (in terms of possible outcomes), and add a framework which alters the ratings on various outcomes in accordance with the model of the preferences. This would be a system which attempts to act as intended; 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. Imagine a human operator who has a friend that must be put into hiding. The operated system may either take the friend to safety, or may 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 direct knowledge of the underlying world-state.

Human preferences are complex, multi-faceted, and often contradictory, and safely extracting preferences from a model of a human is no easy task. Here problems of ontology identification arise again: the framework for extracting preferences and affecting outcome-ratings needs to be robust against drastic changes in the operator-model. 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 (and so on).

DWIM architectures must avoid a number of other hazards, as well. Suppose the system models the fact that its operator-model affects its outcome ratings, and the system has available to it actions which affect the operator. Then actions which manipulate the operator to make their preferences easier to fulfill may 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.

The benefit of a DWIM architecture is that it would allow systems to induce human preferences and act accordingly. Such an architecture requires significant additional complexity on top of inductive value learning: the learning system no longer simply classifies outcomes, it also models humans and extracts human preferences about human-modeled outcomes. What this complexity purchases is a system which potentially achieves full, direct coverage of the complexity of human value, without relying on the abilities of the programmers to hand-code everything or compose the exactly right training-set.

This capability seems critical in the long run, but hard to make immediate research progress upon. Perhaps if a wide space of goal-optimizing procedures is identified, within which learning of particular goal-optimizing procedures is possible, then it might be possible to specify a system that inductively learns how to act as intended. This would be a doubly indirect approach: a hand-coded inductive system would learn from labeled data how to engage in the DWIM procedure that it would use to model operators and act according to its model of the operators' intentions. This would potentially place high demands on corrigibility and the ability to construct systems that behave cautiously in the face of uncertainty.

Further investigations into inverse reinforcement learning (or other methods of constructing satisfactory initial operator-models) may also be a good start on this open problem.

## 5 Extrapolating Volition

A DWIM architecture may be sufficient when constructing a system which 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 (such as "do what I would want, if I had more knowledge and time to think"). None of the goal structures discussed so far seem powerful enough to learn or to capture sophisticated philosophical concepts such as an "ideal advisor theory" (Rosati 1995) or the "reflective equilibrium" of Rawls (1971).

In order to resolve *normative* uncertainty (e.g. about what the operator would want if they were "better"), one possible approach 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 reflecting) adds a new layer of potential pitfalls.[4] 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 begs the question. It is hard for humans to know in advance what sorts of arguments will persuade them, and it seems infeasible to consider all potentially persuasive arguments and categorize them as "valid" or "invalid" forms of persuasion.

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. To expose a human operator to an extremely persuasive moral argument uncovered by superintelligent search, and ask for its label as valid or invalid, may invalidate the resulting label if the argument was in fact invalid but very persuasive. This poses a dilemma for training and validating any system that queries what types of moral arguments should be considered valid persuasion.

A similar set of problems holds for asking a superintelligence to help resolve normatively-laden philosophical questions, such as about what role the notion of consciousness ought to play in value judgments. An obvious but problematic approach would be to task a superintelligent system with producing a philosophical paper on consciousness that a human would find extremely persuasive, or a paper that would make a human feel they had fully understood the problem. Again, in everyday practice, persuasiveness and validity are correlated, but a superintelligent search for an extremely persuasive paper might not pick out a valid one.

The resolution of normative uncertainty seems difficult and potentially dangerous, but it becomes especially important if a superintelligent system is intended to have a large amount of control over the future. This is the motivation for approaches that Bostrom (2014, chap. 13) terms "indirect normativity," by which an agent could learn how to resolve normative uncertainty indirectly.

It is difficult to identify technical approaches to indirect normativity that are tractable today, although some attempts have been made. Christiano (2014) informally proposes one mechanism by which a system could perhaps safely extrapolate the volition of its operator. Fallenstein and Stiennon (2014) have begun examining toy models in which agents operate under uncertainty about which utility function is the "true" utility function; such problems bear a strong resemblance to voter aggregation problems. MacAskill (2014) has given an extensive report on "meta-normativity," touching upon many different philosophical aspects of the difficulties of resolving normative uncertainty. Further philosophical study could lead to progress.

## 6 Discussion

Just as human intelligence has allowed the development of tools and strategies which grant humanity control over the environment, so too could superintelligent systems develop tools and strategies more powerful than our own (Bostrom 2014, chap. 6). It is not clear how long the development of superintelligence will take, and machine superintelligence may not be the first form of superintelligence constructed. But if it is, and if early superintelligent systems are aligned with human interests, then they will likely be controlled by very small groups of humans. Be it one team controlling one superintelligent system or dozens of companies controlling dozens, the development of controllable superintelligence could put the future of humanity into the hands of a shockingly tiny group of people. This introduces a moral hazard of sizable proportions.

If any system is to gain significant control over the future, then it is imperative that the system be constructed to act according to the interests of not only its operators, but all humanity, and perhaps all sapient beings. This, of course, raises yet more questions: How are conflicting preferences resolved? Are children counted? Are animals? Future people? Past people? Again, these are problems of philosophy and are difficult to approach at present. Nevertheless, those who develop the first superintelligent systems take on a sizable responsibility.

That responsibility also demands extreme caution when developing systems intended for superintelligence.

[4] Ethical issues arise when constructing a system that reasons about what its operators would decide if they had more time to think: a sufficiently powerful system might simulate its operators taking the time to think, and a very high-fidelity simulation might be sentient. This concern touches on philosophical questions that are beyond the scope of this paper.

A bug in the value system of a superintelligent agent could be catastrophic, especially if the bug caused the agent to resist correction. Testing alone is not enough: If any important dimensions of value are neglected in the training set, they are likely to also be neglected in the testing environment. Furthermore, a system which models the fact that it is under observation and that its operators do not approve of its preferences, it may well pass all tests, as passing all tests is the only available strategy by which it may exit the testing environment and pursue its actual goals in the world at large. Testing is useful for catching early bugs that occur in settings similar to the test environment; this is essential, but it is not alone enough to gain confidence that the system will work well in reality and in the long run.

Given the enormity of the stakes and the difficulty of writing bug-free software, every available precaution must be taken when constructing superintelligent systems. The system must be corrigible; that is, the structure of its goal system should avert any instrumental incentives to manipulate or deceive its operators, and the system should not resist operator correction or shutdown. Its world model and its decision procedure must be transparent, so that the system may be monitored for hints of manipulation and deception anyway. The system might be "domestic," in the sense of Bostrom (2014, chap. 9), such that its goals would lead it to have only a low impact on the world, if it were to escape. Other safety precautions should be taken, in case clever structuring of the goal system fails to yield a safe system (though failure of the goal system implies that other safety measures may face superintelligent opposition). For further discussion on the design of highly reliable agents, see Soares and Fallenstein (2014).

Value learning is but one component of a safe superintelligent system. That said, successful value learning is of critical importance, for while all other precautions exist to prevent disaster, it is value learning which could enable success.

## References

Bird, Jon, and Paul Layzell. 2002. "The Evolved Radio and Its Implications for Modelling the Evolution of Novel Sensors." In *Proceedings of the 2002 Congress on Evolutionary Computation,* 2:1836–1841. Honolulu, HI: IEEE. doi:10.1109/CEC.2002.1004522.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies.* New York: Oxford University Press.

Christiano, Paul. 2014. "Specifying 'enlightened judgment' precisely (reprise)." *Ordinary Ideas* (blog), August 27. http://ordinaryideas.wordpress.com/2014/08/27/specifying-enlightened-judgment-precisely-reprise/.

De Blanc, Peter. 2011. *Ontological Crises in Artificial Agents' Value Systems.* The Singularity Institute, San Francisco, CA, May 19. http://arxiv.org/abs/1105.3821.

Dreyfus, Hubert L., and Stuart E. Dreyfus. 1992. "What Artificial Experts Can and Cannot Do." *AI & Society* 6 (1): 18–26. doi:10.1007/BF02472766.

Fallenstein, Benja, and Nisan Stiennon. 2014. *"Loudness": On Priors over Preference Relations.* Brief Technical Note. Machine Intelligence Research Institute, Berkeley, CA. http://intelligence.org/files/LoudnessPriors.pdf.

Hibbard, Bill. 2001. "Super-Intelligent Machines." *ACM SIGGRAPH Computer Graphics* 35 (1): 13–15. http://www.siggraph.org/publications/newsletter/issues/v35/v35n1.pdf.

———. 2012. *The Error in My 2001 VisFiles Column,* September. Accessed December 31, 2012. http://www.ssec.wisc.edu/~billh/g/visfiles_error.html.

MacAskill, William. 2014. "Normative Uncertainty." PhD diss., St Anne's College, University of Oxford. http://ora.ox.ac.uk/objects/uuid:8a8b60af-47cd-4abc-9d29-400136c89c0f.

Ng, Andrew Y., and Stuart J. Russell. 2000. "Algorithms for Inverse Reinforcement Learning." In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML-'00),* edited by Pat Langley, 663–670. San Francisco: Morgan Kaufmann.

Rawls, John. 1971. *A Theory of Justice.* Cambridge, MA: Belknap.

Rosati, Connie S. 1995. "Persons, Perspectives, and Full Information Accounts of the Good." *Ethics* 105 (2): 296–325. doi:10.1086/293702.

Russell, Stuart. 2014. "Of Myths and Moonshine." *Edge.org* (blog comment), November. http://edge.org/conversation/the-myth-of-ai#26015.

Schmidhuber, Jürgen. 2007. "Simple Algorithmic Principles of Discovery, Subjective Beauty, Selective Attention, Curiosity and Creativity." In *Discovery Science: 10th International Conference, DS 2007 Sendai, Japan, October 1–4, 2007. Proceedings,* edited by Vincent Corruble, Masayuki Takeda, and Einoshin Suzuki, 26–38. Lecture Notes in Computer Science 4755. Berlin: Springer. doi:10.1007/978-3-540-75488-6_3.

Soares, Nate. 2015. *Formalizing Two Problems of Realistic World-Models.* Technical report 2015–3. Berkeley, CA: Machine Intelligence Research Institute. https://intelligence.org/files/RealisticWorldModels.pdf.

Soares, Nate, and Benja Fallenstein. 2014. *Aligning Superintelligence with Human Interests: A Technical Research Agenda.* Technical report 2014–8. Berkeley, CA: Machine Intelligence Research Institute. https://intelligence.org/files/TechnicalAgenda.pdf.

———. 2015. *Questions of Reasoning Under Logical Uncertainty.* Technical report 2015–1. Berkeley, CA: Machine Intelligence Research Institute. https://intelligence.org/files/QuestionsLogicalUncertainty.pdf.

Sotala, Kaj. Forthcoming. “Concept Learning for Safe Autonomous AI.” Accepted to the 1st International Workshop on AI and Ethics, held within the 29th AAAI Conference on Artificial Intelligence (AAAI-2015), Austin, TX.

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

Yudkowsky, Eliezer. 2011. “Complex Value Systems in Friendly AI.” In *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3–6, 2011. Proceedings,* edited by Jürgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks, 388–393. Lecture Notes in Computer Science 6830. Berlin: Springer. doi:10.1007/978-3-642-22887-2_48.