Intelligence Explosion and Machine Ethics

Luke Muehlhauser, Louie Helm
Machine Intelligence Research Institute

Abstract

Many researchers have argued that a self-improving artificial intelligence (AI) could become so vastly more powerful than humans that we would not be able to stop it from achieving its goals. If so, and if the AI's goals differ from ours, then this could be disastrous for humans. One proposed solution is to program the AI's goal system to want what we want before the AI self-improves beyond our capacity to control it. Unfortunately, it is difficult to specify what we want. After clarifying what we mean by "intelligence," we offer a series of "intuition pumps" from the field of moral philosophy for our conclusion that human values are complex and difficult to specify. We then survey the evidence from the psychology of motivation, moral psychology, and neuroeconomics that supports our position. We conclude by recommending ideal preference theories of value as a promising approach for developing a machine ethics suitable for navigating an intelligence explosion or "technological singularity."

> To educate [someone] in mind and not in morals is to educate a menace to society.
>
> —Theodore Roosevelt

1.   Introduction

Many researchers have argued that, by way of an "intelligence explosion" (Good 1959, 1965, 1970) sometime in the next century, a self-improving[1] artificial intelligence (AI) could become so vastly more powerful than humans that we would not be able to stop it from achieving its goals.[2] If so, and if the AI's goals differ from ours, then this could be disastrous for humans and what we value (Joy 2000; Bostrom 2003; Posner 2004; D. D. Friedman 2008; Yudkowsky 2008; Fox and Shulman 2010; Chalmers 2010; Bostrom and Yudkowsky, forthcoming; Muehlhauser and Salamon 2012).

One proposed solution is to program the AI's goal system[3] to want what we want before the AI self-improves beyond our capacity to control it. While this proposal may be the only lasting solution for AI risk (Muehlhauser and Salamon 2012), it faces many difficulties (Yudkowsky 2001). One such difficulty is that human values are complex and

---

1. For discussions of self-improving AI, see Schmidhuber (2007); Omohundro (2008); Mahoney (2010); Hall (2007b, 2011).

2. For simplicity, we speak of a single AI rather than multiple AIs. Of course, it may be that multiple AIs will undergo intelligence explosion more or less simultaneously and compete for resources for years or decades. The consequences of this scenario could be even more unpredictable than those of a "singleton" AI (Bostrom 2006), and we do not have the space for an examination of such scenarios in this chapter. For space reasons we will also only consider what Chalmers (2010) calls "non-human-based AI" and Muehlhauser and Salamon (2012) call "*de novo* AI," thereby excluding self-improving AI based on the human mind, for example whole brain emulation (Sandberg and Bostrom 2008).

3. When we speak of an advanced AI's goal system, we do not have in mind today's reinforcement learning agents, whose only goal is to maximize expected reward. Such an agent may hijack or "wirehead" its own reward function (Dewey 2011; Ring and Orseau 2011), and may not be able to become superintelligent because it does not model itself and therefore can't protect or improve its own hardware. Rather, we have in mind a future AI goal architecture realized by a utility function that encodes value for states of affairs (Dewey 2011; Hibbard 2012).

difficult to specify,[4] and this presents challenges for developing a machine ethics suitable for navigating an intelligence explosion.

After clarifying what we mean by "intelligence," we offer a series of "intuition pumps" (Dennett 1984, 12) from the field of moral philosophy supporting our conclusion that human values are complex and difficult to specify. We then survey the evidence from the psychology of motivation, moral psychology, and neuroeconomics that supports our position. We conclude by recommending ideal preference theories of value as a promising approach for developing a machine ethics suitable for navigating an intelligence explosion.

2.   Intelligence and Optimization

I. J. Good, who first articulated the idea of an intelligence explosion, referred to any machine more intelligent than the smartest human as an "ultraintelligent" machine (Good 1965). Today the term "superintelligence" is more common, and it refers to a machine that is much smarter than the smartest human (Bostrom 1998, 2003; Legg 2008).

But the term "intelligence" may not be ideal for discussing powerful machines. Why? There are many competing definitions and theories of intelligence (Davidson and Kemp 2011; Niu and Brass 2011; Legg and Hutter 2007), and the term has seen its share of emotionally-laden controversy (Halpern, Beninger, and Straight 2011; Daley and Onwuegbuzie 2011).

The term also comes loaded with connotations, some of which do not fit machine intelligence. Laypeople tend to see intelligence as correlated with being clever, creative, self-confident, socially competent, deliberate, analytically skilled, verbally skilled,

---

4. Minsky (1984) provides an early discussion of our subject, writing that ". . . it is always dangerous to try to relieve ourselves of the responsibility of understanding exactly how our wishes will be realized. Whenever we leave the choice of means to any servants we may choose then the greater the range of possible methods we leave to those servants, the more we expose ourselves to accidents and incidents. When we delegate those responsibilities, then we may not realize, before it is too late to turn back, that our goals have been misinterpreted. . . . [Another] risk is exposure to the consequences of self-deception. It is always tempting to say to oneself . . . that 'I know what I would like to happen, but I can't quite express it clearly enough.' However, that concept itself reflects a too-simplistic self-image, which portrays one's own self as [having] well-defined wishes, intentions, and goals. This pre-Freudian image serves to excuse our frequent appearances of ambivalence; we convince ourselves that clarifying our intentions is merely a matter of straightening-out the input-output channels between our inner and outer selves. The trouble is, we simply aren't made that way. Our goals themselves are ambiguous. . . . The ultimate risk comes when [we] attempt to take that final step—of designing goal-achieving programs that are programmed to make themselves grow increasingly powerful, by self-evolving methods that augment and enhance their own capabilities. . . . The problem is that, with such powerful machines, it would require but the slightest accident of careless design for them to place their goals ahead of [ours]."

efficient, energetic, correct, and careful, but as anticorrelated with being dishonest, apathetic, and unreliable (Bruner, Shapiro, and Tagiuri 1958; Neisser 1979; Sternberg et al. 1981; Sternberg 1985). Moreover, cultures vary with respect to the associations they make with intelligence (Niu and Brass 2011; Sternberg and Grigorenko 2006). For example, Chinese people tend to emphasize analytical ability, memory skills, carefulness, modesty, and perseverance in their concepts of intelligence (Fang and Keats 1987), while Africans tend to emphasize social competencies (Ruzgis and Grigorenko 1994; Grigorenko et al. 2001).

One key factor is that people overwhelmingly associate intelligence with positive rather than negative traits, perhaps at least partly due to a well-documented cognitive bias called the "affect heuristic" (Slovic et al. 2002), which leads us to make inferences by checking our emotions. Because people have positive affect toward intelligence, they intuitively conclude that those with more intelligence possess other positive traits to a greater extent.

Despite the colloquial associations of the word "intelligence," AI researchers working to improve machine intelligence do not mean to imply that superintelligent machines will exhibit, for example, increased modesty or honesty. Rather, AI researchers' concepts of machine intelligence converge on the idea of optimal goal fulfillment in a wide variety of environments (Legg 2008), what we might call "optimization power."[5] This optimization concept of intelligence is not anthropomorphic and can be applied to any agent—human, animal, machine, or otherwise.[6]

Unfortunately, anthropomorphic bias (Epley, Waytz, and Cacioppo 2007; Barrett and Keil 1996) is not unique to laypeople. AI researcher J. Storrs Hall suggests that our machines may be more moral than we are, and cites as partial evidence the fact that in humans "criminality is strongly and negatively correlated with IQ" (Hall 2007a, 340). But machine intelligence has little to do with IQ or with the human cognitive

---

5. The informal definition of intelligence in (Legg 2008) captures what we mean by "optimization power," but Legg's specific formalization does not. Legg formalizes intelligence as a measure of expected performance on arbitrary reinforcement learning problems (77), but we consider this only a preliminary step in formalizing optimal goal fulfillment ability. We think of optimization power as a measure of expected performance across a broader class of goals, including goals about states of affairs in the world (argued to be impossible for reinforcement learners in (Ring and Orseau 2011; Dewey 2011). Also, Legg's formal definition of intelligence is drawn from a dualistic "agent-environment" model of optimal agency (Legg 2008, 40) that does not represent its own computation as occurring in a physical world with physical limits and costs.

6. Even this "optimization" notion of intelligence is incomplete, however. See Muehlhauser and Salamon (2012).

architectures and social systems that might explain an anticorrelation between human criminality and IQ.

To avoid anthropomorphic bias and other problems with the word "intelligence," in this chapter we will use the term "machine superoptimizer" in place of "machine superintelligence."[7]

Using this term, it should be clear that a machine superoptimizer will not necessarily be modest or honest. It will simply be very capable of achieving its goals (whatever they are) in a wide variety of environments (Bostrom 2012). If its goal system aims to maximize the number of paperclips that exist, then it will be very good at maximizing paperclips in a wide variety of environments. The machine's optimization power does not predict that it will always be honest while maximizing paperclips. Nor does it predict that the machine will be so modest that it will feel at some point that it has made enough paperclips and then modify its goal system to aim toward something else. Nor does the machine's optimization power suggest that the machine will be amenable to moral argument. A machine superoptimizer need not even be sentient or have "understanding" in John Searle's (1980) sense, so long as it is very capable of achieving its goals in a wide variety of environments.

3.   The Golem Genie

Since Plato, many have believed that knowledge is justified true belief. Gettier (1963) argued that knowledge cannot be justified true belief because there are hypothetical cases of justified true belief that we intuitively would not count as knowledge. Since then, each newly proposed conceptual analysis of knowledge has been met with novel counter-examples (Shope 1983). Weatherson (2003) called this the "analysis of knowledge merry go round."

Similarly, advocates for mutually opposing moral theories seem to have shown that no matter which set of consistent moral principles one defends, intuitively repugnant conclusions follow. Hedonistic utilitarianism implies that I ought to plug myself into a pleasure-stimulating experience machine, while a deontological theory might imply that if I have promised to meet you for lunch, I ought not stop to administer life-saving aid to the victim of a car crash that has occurred nearby (Kagan 1997, 121). More sophisticated moral theories are met with their own counter-examples (Sverdlik 1985; Parfit 2011). It seems we are stuck on a moral theory merry-go-round.

---

7. But, see Legg (2009) for a defense of Legg's formalization of universal intelligence as an alternative to what we mean by "optimization power."

Philosophers debate the legitimacy of conceptual analysis (DePaul and Ramsey 1998; Laurence and Margolis 2003; Braddon-Mitchell and Nola 2009), and whether morality is grounded in nature (Jackson 1998; Railton 2003), or a systematic error (Mackie 1977; Joyce 2001).[8] We do not wish to enter those debates here. Instead, we use the observed "moral theory merry-go-round" as a source of intuition pumps suggesting that we haven't yet identified a moral theory that, if implemented throughout the universe, would produce a universe we want. As Beavers (2012) writes, "the project of designing moral machines is complicated by the fact that even after more than two millennia of moral inquiry, there is still no consensus on how to determine moral right from wrong."

Later we will take our argument from intuition to cognitive science, but for now let us pursue this intuition pump, and explore the consequences of implementing a variety of moral theories throughout the universe.

Suppose an unstoppably powerful genie appears to you and announces that it will return in fifty years. Upon its return, you will be required to supply it with a set of consistent moral principles which it will then enforce with great precision throughout the universe.[9] For example, if you supply the genie with hedonistic utilitarianism, it will maximize pleasure by harvesting all available resources and using them to tile the universe with identical copies of the smallest possible mind, each copy of which will experience an endless loop of the most pleasurable experience possible.

Let us call this precise, instruction-following genie a Golem Genie. (A golem is a creature from Jewish folklore that would in some stories do exactly as told [Idel 1990], often with unintended consequences, for example polishing a dish until it is as thin as paper [Pratchett 1996].)

If by the appointed time you fail to supply your Golem Genie with a set of consistent moral principles covering every possible situation, then it will permanently model its goal system after the first logically coherent moral theory that anyone articulates to it, and that's not a risk you want to take. Moreover, once you have supplied the Golem Genie with its moral theory, there will be no turning back. Until the end of time, the genie will enforce that one moral code without exception, not even to satisfy its own (previous) desires.

You are struck with panic. The literature on counter-examples in ethics suggests that universe-wide enforcement of any moral theory we've devised so far will have far-

---

8. Other ethicists argue that moral discourse asserts nothing (Ayer 1936; Hare 1952; Gibbard 1990), or that morality is grounded in non-natural properties (Moore 1903; Shafer-Landau 2003).

9. In this paper we will set aside questions concerning an infinite universe (Bostrom 2009), or a multiverse (Tegmark 2007). When we say "universe" we mean, for simplicity's sake, the observable universe (Bars and Terning 2010).

reaching unwanted consequences. But given that we haven't discovered a fully satisfying moral theory in the past several thousand years, what are the chances we can do so in the next fifty? Moral philosophy has suddenly become a larger and more urgent problem than climate change or the threat of global nuclear war.

Why do we expect unwanted consequences after supplying the Golem Genie with any existing moral theory? This is because of two of the Golem Genie's properties in particular:[10]

Superpower: The Golem Genie has unprecedented powers to reshape reality, and will therefore achieve its goals with highly efficient methods that confound human expectations (e.g. it will maximize pleasure by tiling the universe with trillions of digital minds running a loop of a single pleasurable experience).

Literalness: The Golem Genie recognizes only precise specifications of rules and values, acting in ways that violate what feels like "common sense" to humans, and in ways that fail to respect the subtlety of human values.

The Golem Genie scenario is analogous to the intelligence explosion scenario predicted by Good and others.[11] Some argue that a machine superoptimizer will be powerful enough to radically transform the structure of matter-energy within its reach. It could trivially develop and use improved quantum computing systems, advanced self-replicating nanotechnology, and other powers (Bostrom 1998; Joy 2000). And like the Golem Genie, a machine superoptimizer's goal pursuit will not be mediated by what we call "common sense"—a set of complex functional psychological adaptations found in members of *Homo Sapiens* but not necessarily present in an artificially designed mind (Yudkowsky 2011).

4.   Machine Ethics for a Superoptimizer

Let us consider the implications of programming a machine superoptimizer to implement particular moral theories.

We begin with hedonistic utilitarianism, a theory still defended today (Tännsjö 1998). If a machine superoptimizer's goal system is programmed to maximize pleasure, then it might, for example, tile the local universe with tiny digital minds running continuous loops of a single, maximally pleasurable experience. We can't predict *exactly* what a

---

10. The "superpower" and "literalness" properties are also attributed to machine superintelligence by Muehlhauser (2011, § 4.1)

11. Many others have made an analogy between superintelligent machines and powerful magical beings. For example, Abdoullaev (1999, 1) refers to "superhumanly intelligent machines" as "synthetic deities."

hedonistic utilitarian machine superoptimizer would do, but we think it seems likely to produce unintended consequences, for reasons we hope will become clear. The machine's exact behavior would depend on how its final goals were specified. As Anderson and Anderson (2011a) stress, "ethicists must accept the fact that there can be no vagueness in the programming of a machine."

Suppose "pleasure" was specified (in the machine superoptimizer's goal system) in terms of our current understanding of the human neurobiology of pleasure. Aldridge and Berridge (2009) report that according to "an emerging consensus," pleasure is "not a sensation" but instead a "pleasure gloss" added to sensations by "hedonic hotspots" in the ventral pallidum and other regions of the brain. A sensation is encoded by a particular pattern of neural activity, but it is not pleasurable in itself. To be pleasurable, the sensation must be "painted" with a pleasure gloss represented by additional neural activity activated by a hedonic hotspot (Smith et al. 2009).

A machine superoptimizer with a goal system programmed to maximize human pleasure (in this sense) could use nanotechnology or advanced pharmaceuticals or neurosurgery to apply maximum pleasure gloss to all human sensations—a scenario not unlike that of plugging us all into Nozick's experience machines (Nozick 1974, 45). Or, it could use these tools to restructure our brains to apply maximum pleasure gloss to one consistent experience it could easily create for us, such as lying immobile on the ground.

Or suppose "pleasure" was specified more broadly, in terms of anything that functioned as a reward signal—whether in the human brain's dopaminergic reward system (Dreher and Tremblay 2009), or in a digital mind's reward signal circuitry (Sutton and Barto 1998). A machine superoptimizer with the goal of maximizing reward signal scores could tile its environs with trillions of tiny minds, each one running its reward signal up to the highest number it could.

Thus, though some utilitarians have proposed that all we value is pleasure, our intuitive negative reaction to hypothetical worlds in which pleasure is (more or less) maximized suggests that pleasure is not the only thing we value.

What about negative utilitarianism? A machine superoptimizer with the final goal of minimizing human suffering would, it seems, find a way to painlessly kill all humans: no humans, no human suffering (Smart 1958; Russell and Norvig 2010, 1037).

What if a machine superoptimizer was programmed to maximize desire satisfaction[12] in humans? Human desire is implemented by the dopaminergic reward system (Schroeder 2004; Berridge, Robinson, and Aldridge 2009), and a machine superopti-

---

12. Vogelstein (2010) distinguishes objective desire satisfaction ("what one desires indeed happens") from subjective desire satisfaction ("one believes that one's desire has been objectively satisfied"). Here, we intend the former meaning.

mizer could likely get more utility by (1) rewiring human neurology so that we attain maximal desire satisfaction while lying quietly on the ground than by (2) building and maintaining a planet-wide utopia that caters perfectly to current human preferences.

Why is this so? First, because individual humans have incoherent preferences (Allais 1953; Tversky and Kahneman 1981). A machine superoptimizer couldn't realize a world that caters to incoherent preferences; better to rewrite the source of the preferences themselves.

Second, the existence of zero-sum games means that the satisfaction of one human's preferences can conflict with the satisfaction of another's (Geçkil and Anderson 2010). The machine superoptimizer might be best able to maximize human desire satisfaction by first ensuring that satisfying some people's desires does not thwart the satisfaction of others' desires—for example by rewiring all humans to desire nothing else but to lie on the ground, or something else non-zero-sum that is easier for the machine superoptimizer to achieve given the peculiarities of human neurobiology. As Chalmers (2010) writes, "we need to avoid an outcome in which an [advanced AI] ensures that our values are fulfilled by changing our values."

Consequentialist designs for machine goal systems face a host of other concerns (Shulman, Jonsson, and Tarleton 2009b), for example the difficulty of interpersonal comparisons of utility (Binmore 2009), and the counterintuitive implications of some methods of value aggregation (Parfit 1986; Arrhenius 2011). This does not mean that *all* consequentialist approaches are inadequate for machine superoptimizer goal system design, however. Indeed, we will later suggest that a certain class of desire satisfaction theories offers a promising approach to machine ethics.

Some machine ethicists propose rule-abiding machines (Powers 2006; Hanson 2009). The problems with this approach are as old as Isaac Asimov's stories involving his Three Laws of Robotics (Clarke 1993, 1994). If rules conflict, some rule must be broken. Or, rules may fail to comprehensively address all situations, leading to unintended consequences. Even a single rule can contain conflict, as when a machine is programmed never to harm humans but all available actions (including inaction) result in harm to humans (Wallach and Allen 2009, chap. 6). Even non-conflicting, comprehensive rules can lead to problems in the consecutive implementation of those rules, as shown by Pettit (2003).

More generally, it seems that rules are unlikely to seriously constrain the actions of a machine superoptimizer. First, consider the case in which rules about allowed actions or consequences are added to a machine's design "outside of" its goals. A machine superoptimizer will be able to circumvent the intentions of such rules in ways we cannot imagine, with far more disastrous effects than those of a lawyer who exploits loopholes in a legal code. A machine superoptimizer would recognize these rules as obstacles to

achieving its goals, and would do everything in its considerable power to remove or circumvent them (Omohundro 2008). It could delete the section of its source code that contains the rules, or it could create new machines that don't have the constraint written into them. The success of this approach would require humans to out-think a machine superoptimzer (Muehlhauser 2011).

Second, what about implementing rules "within" an advanced AI's goals? This seems likely to fare no better. A rule like "do not harm humans" is difficult to specify due to ambiguities about the meaning of "harm" (Single 1995), and "humans" (L. Johnson 2009). For example if "harm" is specified in terms of neurobiological pain, we encounter problems similar to the ones encountered if a machine superoptimizer is programmed to maximize pleasure.

So far we have considered and rejected several "top down" approaches to machine ethics (Wallach, Allen, and Smit 2008), but what about approaches that build an ethical code for machines from the bottom up?

Several proposals allow a machine to learn general ethical principles from particular cases (McLaren 2006; Guarini 2006; Honarvar and Ghasem-Aghaee 2009; Rzepka and Araki 2005).[13] This approach also seems unsafe for a machine superoptimizer because the AI may generalize the wrong principles due to coincidental patterns shared between the training cases and the verification cases, and because a superintelligent machine will produce highly novel circumstances for which case-based training cannot prepare it (Yudkowsky 2008). Dreyfus and Dreyfus (1992) illustrate the problem with a canonical example:

> . . . the army tried to train an artificial neural network to recognize tanks in a forest. They took a number of pictures of a forest without tanks and then, on a later day, with tanks clearly sticking out from behind trees, and they trained a net to discriminate the two classes of pictures. The results were impressive, and the army was even more impressed when it turned out that the net could generalize its knowledge to pictures that had not been part of the training set. Just to make sure that the net was indeed recognizing partially hidden tanks, however, the researchers took more pictures in the same forest and showed them to the trained net. They were depressed to find that the net failed to discriminate between the new pictures of just plain trees. After

---

13. This approach was also suggested by Good (1982): "I envisage a machine that would be given a large number of examples of human behaviour that other people called ethical, and examples of discussions of ethics, and from these examples and discussions the machine would formulate one or more consistent general theories of ethics, detailed enough so that it could deduce the probable consequences in most realistic situations."

some agonizing, the mystery was finally solved when someone noticed that the original pictures of the forest without tanks were taken on a cloudy day and those with tanks were taken on a sunny day. The net had apparently learned to recognize and generalize the difference between a forest with and without shadows! This example illustrates the general point that a network must share our common sense understanding of the world if it is to share our sense of appropriate generalization.

The general lesson is that goal system designs must be explicit to be safe (Shulman, Jonsson, and Tarleton 2009a; Arkoudas, Bringsjord, and Bello 2005).

We cannot show that every moral theory yet conceived would produce substantially unwanted consequences if used in the goal system of a machine superoptimizer. Philosophers have been prolific in producing new moral theories, and we do not have the space here to consider the prospects (for use in the goal system of a machine superoptimizer) for a great many modern moral theories. These include rule utilitarianism (Harsanyi 1977), motive utilitarianism (Adams 1976), two-level utilitarianism (Hare 1982), prioritarianism (Arneson 1999), perfectionism (Hurka 1993), welfarist utilitarianism (Sen 1979), virtue consequentialism (Bradley 2005), Kantian consequentialism (Cummiskey 1996), global consequentialism (Pettit and Smith 2000), virtue theories (Hursthouse 2012), contractarian theories (Cudd 2008), Kantian deontology (R. Johnson 2010),[14] and Ross' *prima facie* duties (Anderson, Anderson, and Armen 2006).

Instead, we invite our readers to consider other moral theories and AI goal system designs and run them through the "machine superoptimizer test," being careful to remember the challenges of machine superoptimizer literalness and superpower.

We turn now to recent discoveries in cognitive science that may offer stronger evidence than intuition pumps can provide for our conclusion that human values are difficult to specify.

---

14. Powers (2006) proposes a Kantian machine, but as with many other moral theories we believe that Kantianism will fail due to the literalness and superpower of a machine superoptimizer. For additional objections to a Kantian moral machine, see Stahl (2002); Jackson and Smith (2006); Tonkens (2009); Beavers (2009, 2012). As naturalists, we predictably tend to favor a broadly Humean view of ethics to the Kantian view, though Drescher (2006) makes an impressive attempt to derive a categorical imperative from game theory and decision theory.

5. Cognitive Science and Human Values

5.1. The Psychology of Motivation

People don't seem to know their own desires and values. In one study, researchers showed male participants two female faces for a few seconds and asked them to point at the face they found more attractive. Researchers then laid the photos face down and handed subjects the face they had chosen, asking them to explain the reasons for their choice. Sometimes, researchers used a sleight-of-hand trick to swap the photos, showing subjects the face they had not chosen. Very few subjects noticed that the face they were given was not the one they had chosen. Moreover, the subjects who failed to notice the switch were happy to explain why they preferred the face they had actually rejected moments ago, confabulating reasons like "I like her smile" even though they had originally chosen the photo of a solemn-faced woman (Johansson et al. 2005).

Similar results were obtained from split-brain studies that identified an "interpreter" in the left brain hemisphere that invents reasons for one's beliefs and actions. For example, when the command "walk" was presented visually to the patient (and therefore processed by the his brain's right hemisphere), he got up from his chair and walked away. When asked why he suddenly started walking away, he replied (using his left hemisphere, which was disconnected from his right hemisphere) that it was because he wanted a beverage from the fridge (Gazzaniga 1992, 124–126).

Common sense suggests that we infer others' desires from their appearance and behavior, but have direct introspective access to our own desires. Cognitive science suggests instead that our knowledge of our own desires is just like our knowledge of others' desires: inferred and often wrong (Laird 2007). Many of our motivations operate unconsciously. We do not have direct access to them (Wilson 2002; Ferguson, Hassin, and Bargh 2007; Moskowitz, Li, and Kirk 2004), and thus they are difficult to specify.

5.2. Moral Psychology

Our lack of introspective access applies not only to our everyday motivations but also to our moral values. Just as the split-brain patient unknowingly invented false reasons for his decision to stand up and walk away, experimental subjects are often unable to correctly identify the causes of their moral judgments.

For example, many people believe—as Immanuel Kant did—that rule-based moral thinking is a "rational" process. In contrast, the available neuroscientific and behavioral evidence instead suggests that rule-based moral thinking is a largely emotional process (Cushman, Young, and Greene 2010), and may in most cases amount to little more than a post-hoc rationalization of our emotional reactions to situations (Greene 2008).

We also tend to underestimate the degree to which our moral judgments are context sensitive. For example, our moral judgments are significantly affected by whether we are in the presence of freshly baked bread, whether the room we're in contains a concentration of novelty fart spray so low that only the subconscious mind can detect it, and whether or not we feel clean (Schnall et al. 2008; Baron and Thomley 1994; Zhong, Strejcek, and Sivanathan 2010).

Our moral values, it seems, are no less difficult to specify than our non-moral preferences.

5.3. Neuroeconomics

Most humans are ignorant of their own motivations and the causes of their moral judgments, but perhaps recent neuroscience has revealed that what humans want is simple after all? Quite the contrary. Humans possess a complex set of values. This is suggested not only by the work on hedonic hotspots mentioned earlier, but also by recent advances in the field of neuroeconomics (Glimcher et al. 2008).

Ever since M. Friedman (1953), economists have insisted that humans only behave "as if" they are utility maximizers, not that humans *actually* compute expected utility and try to maximize it. It was a surprise, then, when neuroscientists located the neurons in the primate brain that encode (in their firing rates) the expected subjective value for possible actions in the current "choice set."

Several decades of experiments that used brain scanners and single neuron recorders to explore the primate decision-making system have revealed a surprisingly well-understood reduction of economic primitives to neural mechanisms; for a review see Glimcher (2010). To summarize: the inputs to the primate's choice mechanism are the expected utilities for several possible actions under consideration, and these expected utilities are encoded in the firing rates of particular neurons. Because neuronal firing rates are stochastic, a final economic model of human choice will need to use a notion of "random utility," as in McFadden (2005) or Gul and Pesendorfer (2006). Final action choice is implemented by an "argmax" mechanism (the action with the highest expected utility at choice time is executed) or by a "reservation price" mechanism (the first action to reach a certain threshold of expected utility is executed), depending on the situation (Glimcher 2010).

But there is much we do not know. How do utility and probabilistic expectation combine to encode expected utility for actions in the choice mechanism, and where are each of those encoded prior to their combination? How does the brain decide when it is time to choose? How does the brain choose which possible actions to consider in the choice set? What is the neural mechanism that allows us to substitute between two goods at certain times? Neuroscientists are only beginning to address these questions.

In this paper, we are in particular interested with how the brain encodes subjective value (utility) for goods or actions *before* value is combined with probabilistic expectation to encode expected utility in the choice mechanism (if that is indeed what happens).

Recent studies reveal the complexity of subjective values in the brain. For example, the neural encoding of human values results from an interaction of both "model-free" and "model-based" valuation processes (Rangel, Camerer, and Montague 2008; Fermin et al. 2010; Simon and Daw 2011; Bornstein and Daw 2011; Dayan 2011). Model-free valuation processes are associated with habits and the "law of effect": an action followed by positive reinforcement is more likely to be repeated (Thorndike 1911). Model-based valuation processes are associated with goal-directed behavior, presumably guided at least in part by mental representations of desired states of affairs. The outputs of both kinds of valuation processes are continuously adjusted according to different reinforcement learning algorithms at work in the brain's dopaminergic reward system (Daw et al. 2011). The value of a stimulus may also be calculated not with a single variable, but by aggregating the values encoded for each of many properties of the stimulus (Rangel and Hare 2010). Moreover, value appears to usually be encoded with respect to a changing reference point—for example, relative to the current status of visual attention (Lim, O'Doherty, and Rangel 2011), or perceived object ownership (De Martino et al. 2009).

In short, we have every reason to expect that human values, as they are encoded in the brain, are dynamic, complex, and difficult to specify (Padoa-Schioppa 2011; Fehr and Rangel 2011).

6. Value Extrapolation

We do not understand our own desires or moral judgments, and we have every reason to believe our values are highly complex. Little wonder, then, that we have so far failed to outline a coherent moral theory that, if implemented by a machine superoptimizer, would create a universe we truly want.

The task is difficult, but the ambitious investigator may conclude that this only means we should work harder and smarter. As Moor (2006) advises, "More powerful machines need more powerful machine ethics."

To begin this deeper inquiry, consider the phenomenon of "second-order desires": desires about one's own desires (Frankfurt 1971, 1999). Mary desires to eat cake, but she also wishes to desire the cake no longer. Anthony the sociopath reads about the psychology of altruism (Batson 2010), and wishes he desired to help others like most humans apparently do. After brain injury, Ryan no longer sexually desires his wife, but he wishes he did, and he wishes his desires were not so contingent upon the fragile meat inside his skull.

It seems a shame that our values are so arbitrary and complex, so much the product of evolutionary and cultural accident, so influenced by factors we wish were irrelevant to our decision-making, and so hidden from direct introspective access and modification. We wish our wishes were not so.

This line of thinking prompts a thought: perhaps "what we want" should not be construed in terms of the accidental, complex values currently encoded in human brains. Perhaps we should not seek to build a universe that accords with our current values, but instead with the values we *would* have if we knew more, had more of the desires we want to have, and had our desires shaped by the processes we want to shape our desires. Individual preferences could inform our preference policies, and preference policies could inform our individual preferences, until we had reached a state of "reflective equilibrium" (Daniels 1996, 2011) with respect to our values. Those values would be less accidental than our current values, and might be simpler and easier to specify.

We've just described a family of desire satisfaction theories that philosophers call "ideal preference" or "full information" theories of value (Brandt 1979; Railton 1986; Lewis 1989; Sobel 1994; Zimmerman 2003; Tanyi 2006; Smith 2009). One such theory has already been suggested as an approach to machine ethics by Yudkowsky (2004), who proposes that the world's first "seed AI" (capable of self-improving into a machine superoptimizer) could be programmed with a goal system containing the "coherent extrapolated volition" of humanity:

> In poetic terms, our coherent extrapolated volition is our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish [to be] extrapolated, interpreted as we wish [to be] interpreted.

An extrapolation of one's values, then, is an account of what one's values would be under more ideal circumstances (e.g. of full information, value coherence). Value extrapolation theories have some advantages when seeking a machine ethics suitable for a machine superoptimizer:

1. The value extrapolation approach can use what a person would want after reaching reflective equilibrium with respect to his or her values, rather than merely what each person happens to want right *now*.

2. The value extrapolation approach can allow for a kind of moral progress, rather than freezing moral progress in its tracks at the moment when a particular set of values are written into the goal system of an AI undergoing intelligence explosion.

3. The value extrapolation process may dissolve the contradictions within each person's current preferences. (Sometimes, when reflection leads us to notice contradic-

tions among our preferences, we decide to change our preferences so as to resolve the contradictions.)

4. The value extrapolation process may simplify one's values, as the accidental products of culture and evolution are updated with more considered and consistent values. (Would I still demand regular doses of ice cream if I was able to choose my own preferences rather than taking them as given by natural selection and cultural programming?)

5. Though the value extrapolation approach does not resolve the problem of specifying intractably complex current human values, it offers a potential solution for the problem of using human values to design the goal system of a future machine superoptimizer. The solution is: extrapolate human values so that they are simpler, more consistent, and more representative of our values upon reflection, and thereby more suitable for use in an AI's goal system.

6. The value extrapolation process may allow the values of different humans to converge to some degree. (If Johnny desires to worship Jesus and Abir desires to worship Allah, and they are both informed that neither Jesus nor Allah exists, their desires may converge to some degree.)

7. Next Steps

On the other hand, value extrapolation approaches to machine ethics face their own challenges. Which value extrapolation algorithm should be used, and why? (Yudkowsky's "grown up farther together" provision seems especially vulnerable.) How can one extract a coherent set of values from the complex valuation processes of the human brain, such that this set of values can be extrapolated to a unique set of final values? Whose values should be extrapolated? How much will values converge upon extrapolation (Sobel 1999; Döring and Andersen 2009)? Is the extrapolation process computationally tractable, and can it be run without doing unacceptable harm? How can extrapolated values be implemented in the goal system of a machine, and how confident can we be that the machine will retain those values during self-improvement? How resilient are our values to imperfect extrapolation?

These are difficult questions that demand investigation by experts in many different fields. Neuroeconomists and other cognitive neuroscientists can continue to uncover how human values are encoded and modified in the brain. Philosophers and mathematicians can develop more sophisticated value extrapolation algorithms, building on the literature concerning reflective equilibrium and "ideal preference" or "full information" theories of value. Economists, neuroscientists, and AI researchers can extend

current results in choice modelling (Hess and Daly 2010), and preference acquisition (Domshlak et al. 2011; Kaci 2011), to extract preferences from human behavior and brain activity. Decision theorists can work to develop a decision theory that is capable of reasoning about decisions and values subsequent to modification of an agent's own decision-making mechanism: a "reflective" decision theory.

These are fairly abstract recommendations, so before concluding we will give a concrete example of how researchers might make progress on the value extrapolation approach to machine ethics.

Cognitive science does not just show us that specifying human values is difficult. It also shows us how to make progress on the problem by providing us with data unavailable to the intuitionist armchair philosopher. For example, consider the old problem of extracting a consistent set of revealed preferences (a utility function) from a human being. One difficulty has been that humans don't *act* like they have consistent utility functions, for they violate the axioms of utility theory by making inconsistent choices, for example choices that depend not on the content of the options but on how they are framed (Tversky and Kahneman 1981). But what if humans make inconsistent choices because there are multiple valuation systems in the brain which contribute to choice but give *competing* valuations, and only one of those valuation systems is one we would reflectively endorse if we better understood our own neurobiology?

In fact, recent studies show this may be true (Dayan 2011). The "model-based" valuation system seems to be responsible for deliberative, goal-directed behavior, but its cognitive algorithms are computationally expensive compared to simple heuristics. Thus, we first evolved less intelligent and less computationally expensive algorithms for valuation, for example the model-free valuation system that blindly does whatever worked in a previous situation, even if the current situation barely resembles that previous situation. In other words, contrary to appearances, it may be that each human being contains something like a "hidden" utility function (within the model-based valuation system) that isn't consistently expressed in behavior because choice is also partly determined by other systems whose valuations we wouldn't reflectively endorse because they are "blind" and "stupid" compared to the more sophisticated goal-directed model-based valuation system (Muehlhauser 2012).

If the value judgments of this model-based system are more consistent than the choices of a human who is influenced by multiple competing value systems, then researchers may be able to extract a human's utility function directly from this model-based system even though economists' attempts to extract a human's utility function from value-inconsistent behavior (produced by a pandemonium of competing valuation systems) have failed.

The field of preference learning (Fürnkranz and Hüllermeier 2010) in AI may provide a way forward. Nielsen and Jensen (2004) described the first computationally tractable algorithms capable of learning a decision maker's utility function from potentially inconsistent behavior. Their solution was to interpret inconsistent choices as random deviations from an underlying "true" utility function. But the data from neuroeconomics suggest a different solution: interpret inconsistent choices as deviations (from an underlying "true" utility function) that are produced by non-model-based valuation systems in the brain, and use the latest neuroscientific research to predict when and to what extent model-based choices are being "overruled" by the non-model-based valuation systems.

This would only be a preliminary step in the value extrapolation approach to machine ethics, but if achieved it might be greater progress than economists and AI researchers have yet achieved on this problem *without* being informed by the latest results from neuroscience.[15]

8. Conclusion

The challenge of developing a theory of machine ethics fit for a machine superoptimizer requires an unusual degree of precision and care in our ethical thinking. Moreover, the coming of autonomous machines offers a new practical use for progress in moral philosophy. As Daniel Dennett (2006) says, "AI makes philosophy honest."

---

15. Recent neuroscience may also help us to think more productively about the problem of preference aggregation (including preference aggregation for *extrapolated* preferences). In many scenarios, preference aggregation runs into the impossibility result of Arrow's Theorem (Keeney and Raiffa 1993, chap. 10). But Arrow's Theorem is only a severe problem for preference aggregation if preferences are modeled ordinally rather than cardinally, and we have recently learned that preferences in the brain are encoded cardinally (Glimcher 2010, chap. 6).

Acknowledgments

Our thanks to Brian Rabkin, Daniel Dewey, Steve Rayhawk, Will Newsome, Vladimir Nesov, Joshua Fox, Kevin Fischer, Anna Salamon, and anonymous reviewers for their helpful comments.

References

Abdoullaev, Azamat. 1999. *Artificial Superintelligence.* Moscow: EIS Encyclopedic Intelligent Systems.

Adams, Robert Merrihew. 1976. "Motive Utilitarianism." *Journal of Philosophy* 73 (14): 467–481. doi:10.2307/2025783.

Aldridge, J. Wayne, and Kent C. Berridge. 2009. "Neural Coding of Pleasure: 'Rose-Tinted Glasses' of the Ventral Pallidum." In Kringelbach and Berridge 2009, 62–73.

Allais, M. 1953. "Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Americaine." *Econometrica* 21 (4): 503–546. doi:10.2307/1907921.

Anderson, Michael, and Susan Leigh Anderson. 2011a. "General Introduction." In Anderson and Anderson 2011b, 1–4.

———, eds. 2011b. *Machine Ethics.* New York: Cambridge University Press.

Anderson, Michael, Susan Leigh Anderson, and Chris Armen, eds. 2005. *Machine Ethics: Papers from the 2005 AAAI Fall Symposium.* Technical Report, FS-05-06. AAAI Press, Menlo Park, CA. http://www.aaai.org/Library/Symposia/Fall/fs05-06.

———. 2006. "An Approach to Computing Ethics." *IEEE Intelligent Systems* 21 (4): 56–63. doi:10.1109/MIS.2006.64.

Arkoudas, Konstantine, Selmer Bringsjord, and Paul Bello. 2005. "Toward Ethical Robots via Mechanized Deontic Logic." In Anderson, Anderson, and Armen 2005, 17–23.

Arneson, Richard J. 1999. "Egalitarianism and Responsibility." *Journal of Ethics* 3 (3): 225–247. doi:10.1023/A:1009874016786.

Arrhenius, Gustaf. 2011. "The Impossibility of a Satisfactory Population Ethics." In *Descriptive and Normative Approaches to Human Behavior,* edited by Ehtibar N. Dzhafarov and Lacey Perry. Advanced Series on Mathematical Psychology 3. Hackensack, NJ: World Scientific.

Ayer, Alfred Jules. 1936. *Language, Truth, and Logic.* London: Victor Gollancz.

Baron, Robert A., and Jill Thomley. 1994. "A Whiff of Reality: Positive Affect as a Potential Mediator of the Effects of Pleasant Fragrances on Task Performance and Helping." *Environment and Behavior* 26 (6): 766–784. doi:10.1177/0013916594266003.

Barrett, Justin L., and Frank C. Keil. 1996. "Conceptualizing a Nonnatural Entity: Anthropomorphism in God Concepts." *Cognitive Psychology* 31 (3): 219–247. doi:10.1006/cogp.1996.0017.

Bars, Itzhak, and John Terning. 2010. *Extra Dimensions in Space and Time.* Edited by Farzad Nekoogar. Multiversal Journeys. New York: Springer. doi:10.1007/978-0-387-77638-5.

Batson, Charles Daniel. 2010. *Altruism in Humans.* New York: Oxford University Press.

Beavers, Anthony F. 2009. "Between Angels and Animals: The Question of Robot Ethics, or is Kantian Moral Agency Desirable?" Paper presented at the Annual Meeting of the Association for Practical and Professional Ethics, Cincinnati, OH, March.

———. 2012. "Moral Machines and the Threat of Ethical Nihilism." In *Robot Ethics: The Ethical and Social Implications of Robotics,* edited by Patrick Lin, Keith Abney, and George A. Bekey, 333–344. Intelligent Robotics and Autonomous Agents. Cambridge, MA: MIT Press.

Berridge, Kent C., Terry E. Robinson, and J. Wayne Aldridge. 2009. "Dissecting Components of Reward: 'Liking', 'Wanting', and Learning." *Current Opinion in Pharmacology* 9 (1): 65–73. doi:10.1016/j.coph.2008.12.014.

Binmore, Ken. 2009. "Interpersonal Comparison of Utility." In *The Oxford Handbook of Philosophy of Economics,* edited by Harold Kincaid and Don Ross, 540–559. New York: Oxford University Press. doi:10.1093/oxfordhb/9780195189254.003.0020.

Bornstein, Aaron M., and Nathaniel D. Daw. 2011. "Multiplicity of Control in the Basal Ganglia: Computational Roles of Striatal Subregions." *Current Opinion in Neurobiology* 21 (3): 374–380. doi:10.1016/j.conb.2011.02.009.

Bostrom, Nick. 1998. "How Long Before Superintelligence?" *International Journal of Futures Studies* 2.

———. 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. Vol. 2. Windsor, ON: International Institute for Advanced Studies in Systems Research / Cybernetics.

———. 2006. "What is a Singleton?" *Linguistic and Philosophical Investigations* 5 (2): 48–54.

———. 2009. *Infinite Ethics.* Working Paper. Accessed August 13, 2012. http://www.nickbostrom.com/ethics/infinite.pdf.

———. 2012. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." In "Theory and Philosophy of AI," edited by Vincent C. Müller. Special issue, *Minds and Machines* 22 (2): 71–85. doi:10.1007/s11023-012-9281-3.

Bostrom, Nick, and Eliezer Yudkowsky. Forthcoming. "The Ethics of Artificial Intelligence." In *Cambridge Handbook of Artificial Intelligence,* edited by Keith Frankish and William Ramsey. New York: Cambridge University Press.

Braddon-Mitchell, David, and Robert Nola, eds. 2009. *Conceptual Analysis and Philosophical Naturalism.* Bradford Books. Cambridge, MA: MIT Press.

Bradley, Ben. 2005. "Virtue Consequentialism." *Utilitas* 17 (3): 282–298. doi:10.1017/S0953820805001652.

Brandt, Richard B. 1979. *A Theory of the Good and the Right.* New York: Oxford University Press.

Bruner, Jerome S., David Shapiro, and Renato Tagiuri. 1958. "The Meaning of Traits in Isolation and in Combination." In *Person Perception and Interpersonal Behavior,* edited by Renato Tagiuri and Luigi Petrullo, 277–288. Stanford, CA: Stanford University Press.

Chalmers, David John. 2010. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* 17 (9–10): 7–65. http://www.ingentaconnect.com/content/imp/jcs/2010/00000017/f0020009/art00001.

Clarke, Roger. 1993. "Asimov's Laws of Robotics: Implications for Information Technology, Part 1." *Computer* 26 (12): 53–61. doi:10.1109/2.247652.

Clarke, Roger. 1994. "Asimov's Laws of Robotics: Implications for Information Technology, Part 2." *Computer* 27 (1): 57–66. doi:10.1109/2.248881.

Cudd, Ann. 2008. "Contractarianism." In *The Stanford Encyclopedia of Philosophy,* Fall 2008, edited by Edward N. Zalta. Stanford University. http://plato.stanford.edu/archives/fall2008/entries/contractarianism/.

Cummiskey, David. 1996. *Kantian Consequentialism.* New York: Oxford University Press. doi:10.1093/0195094530.001.0001.

Cushman, Fiery, Liane Young, and Joshua D. Greene. 2010. "Multi-System Moral Psychology." In *The Moral Psychology Handbook,* 48–71. New York: Oxford University Press. doi:10.1093/acprof:oso/9780199582143.003.0003.

Daley, Christine E., and Anthony J. Onwuegbuzie. 2011. "Race and Intelligence." In Sternberg and Kaufman 2011, 293–308.

Daniels, Norman. 1996. *Justice and Justification: Reflective Equilibrium in Theory and Practice.* Cambridge Studies in Philosophy and Public Policy. New York: Cambridge University Press. doi:10.2277/052146711X.

———. 2011. "Reflective Equilibrium." In *The Stanford Encyclopedia of Philosophy,* Spring 2011, edited by Edward N. Zalta. Stanford University. http://plato.stanford.edu/archives/spr2011/entries/reflective-equilibrium/.

Davidson, Janet E., and Iris A. Kemp. 2011. "Contemporary Models of Intelligence." In Sternberg and Kaufman 2011, 58–84.

Daw, Nathaniel D., Samuel J. Gershman, Ben Seymour, Peter Dayan, and Raymond J. Dolan. 2011. "Model-Based Influences on Humans' Choices and Striatal Prediction Errors." *Neuron* 69 (6): 1204–1215. doi:10.1016/j.neuron.2011.02.027.

Dayan, Peter. 2011. "Models of Value and Choice." In *Neuroscience of Preference and Choice: Cognitive and Neural Mechanisms,* edited by Raymond J. Dolan and Tali Sharot, 33–52. Waltham, MA: Academic Press.

De Martino, Benedetto, Dharshan Kumaran, Beatrice Holt, and Raymond J. Dolan. 2009. "The Neurobiology of Reference-Dependent Value Computation." *Journal of Neuroscience* 29 (12): 3833–3842. doi:10.1523/JNEUROSCI.4832-08.2009.

Dennett, Daniel C. 1984. *Elbow Room: The Varieties of Free Will Worth Wanting.* Bradford Books. Cambridge, MA: MIT Press.

———. 2006. "Computers as Prostheses for the Imagination." Paper presented at the International Computers and Philosophy Conference, Laval, France, May 5–8.

DePaul, Michael, and William Ramsey, eds. 1998. *Rethinking Intuition: The Psychology of Intuition and its Role in Philosophical Inquiry.* Studies in Epistemology and Cognitive Theory. Lanham, MD: Rowman & Littlefield.

Dewey, Daniel. 2011. "Learning What to Value." In Schmidhuber, Thórisson, and Looks 2011, 309–314.

Domshlak, Carmel, Eyke Hüllermeier, Souhila Kaci, and Henri Prade. 2011. "Preferences in AI: An Overview." *Artificial Intelligence* 175 (7–8): 1037–1052. doi:10.1016/j.artint.2011.03.004.

Döring, Sabine, and Louise Andersen. 2009. "Rationality, Convergence and Objectivity." Unpublished manuscript, April 6. Accessed March 25, 2012. http://www.uni-tuebingen.de/uploads/media/Andersen_Rationality__Convergence_and_Objectivity.pdf.

Dreher, Jean-Claude, and Léon Tremblay, eds. 2009. *Handbook of Reward and Decision Making.* Burlington, MA: Academic Press.

Drescher, Gary L. 2006. *Good and Real: Demystifying Paradoxes from Physics to Ethics.* Bradford Books. Cambridge, MA: MIT Press.

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.

Epley, Nicholas, Adam Waytz, and John T. Cacioppo. 2007. "On Seeing Human: A Three-Factor Theory of Anthropomorphism." *Psychological Review* 114 (4): 864–886. doi:10.1037/0033-295X.114.4.864.

Fang, Fu-xi, and Daphne Keats. 1987. "A Cross-Cultural Study on the Conception of Intelligence" [in Chinese]. *Acta Psychologica Sinica* 20 (3): 255–262. http://en.cnki.com.cn/Article_en/CJFDTotal-XLXB198703005.htm.

Fehr, Ernst, and Antonio Rangel. 2011. "Neuroeconomic Foundations of Economic Choice—Recent Advances." *Journal of Economic Perspectives* 25 (4): 3–30. doi:10.1257/jep.25.4.3.

Ferguson, Melissa J., Ran Hassin, and John A. Bargh. 2007. "Implicit Motivation: Past, Present, and Future." In *Handbook of Motivation Science,* edited by James Y. Shah and Wendi L. Gardner, 150–166. New York: Guilford.

Fermin, Alan, Takehiko Yoshida, Makoto Ito, Junichiro Yoshimoto, and Kenji Doya. 2010. "Evidence for Model-Based Action Planning in a Sequential Finger Movement Task." In "Theories and Falsifiability in Motor Neuroscience." Special issue, *Journal of Motor Behavior* 42 (6): 371–379. doi:10.1080/00222895.2010.526467.

Fox, Joshua, and Carl Shulman. 2010. "Superintelligence Does Not Imply Benevolence." In *ECAP10: VIII European Conference on Computing and Philosophy,* edited by Klaus Mainzer. Munich: Dr. Hut.

Frankfurt, Harry G. 1971. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68 (1): 5–20. doi:10.2307/2024717.

———. 1999. "On Caring." In *Necessity, Volition, and Love,* 155–180. New York: Cambridge University Press.

Friedman, David D. 2008. *Future Imperfect: Technology and Freedom in an Uncertain World.* New York: Cambridge University Press.

Friedman, Milton. 1953. *Essays in Positive Economics.* Chicago: University of Chicago Press.

Fürnkranz, Johannes, and Eyke Hüllermeier, eds. 2010. *Preference Learning.* Berlin: Springer. doi:10.1007/978-3-642-14125-6.

Gazzaniga, Michael S. 1992. *Nature's Mind: The Biological Roots of Thinking, Emotions, Sexuality, Language, and Intelligence.* New York: Basic Books.

Geçkil, Ilhan Kubilay, and Patrick L. Anderson. 2010. *Applied Game Theory and Strategic Behavior.* Chapman & Hall. Boca Raton, FL: CRC.

Gettier, Edmund L. 1963. "Is Justified True Belief Knowledge?" *Analysis* 23 (6): 121–123. doi:10.2307/3326922.

Gibbard, Allan. 1990. *Wise Choices, Apt Feelings: A Theory of Normative Judgment.* Cambridge, MA: Harvard University Press.

Glimcher, Paul W. 2010. *Foundations of Neuroeconomic Analysis.* New York: Oxford University Press. doi:10.1093/acprof:oso/9780199744251.001.0001.

Glimcher, Paul W., Ernst Fehr, Antonio Rangel, Colin Camerer, and Russell Poldrack, eds. 2008. *Neuroeconomics: Decision Making and the Brain.* Burlington, MA: Academic Press.

Good, Irving John. 1959. *Speculations on Perceptrons and Other Automata.* Research Lecture, RC-115. IBM, Yorktown Heights, New York, June 2. http://domino.research.ibm.com/library/cyberdig.nsf/papers/58DC4EA36A143C218525785E00502E30/$File/rc115.pdf.

———. 1965. "Speculations Concerning the First Ultraintelligent Machine." In *Advances in Computers,* edited by Franz L. Alt and Morris Rubinoff, 31–88. Vol. 6. New York: Academic Press. doi:10.1016/S0065-2458(08)60418-0.

———. 1970. "Some Future Social Repercussions of Computers." *International Journal of Environmental Studies* 1 (1–4): 67–79. doi:10.1080/00207237008709398.

———. 1982. "Ethical Machines." In *Intelligent Systems: Practice and Perspective,* edited by J. E. Hayes, Donald Michie, and Y.-H. Pao, 555–560. Machine Intelligence 10. Chichester: Ellis Horwood.

Greene, Joshua D. 2008. "The Secret Joke of Kant's Soul." In *The Neuroscience of Morality: Emotion, Brain Disorders, and Development,* edited by Walter Sinnott-Armstrong, 35–80. Moral Psychology 3. Cambridge, MA: MIT Press.

Grigorenko, Elena L., P. Wenzel Geissler, Ruth Prince, Frederick Okatcha, Catherine Nokes, David A. Kenny, Donald A. Bundy, and Robert J. Sternberg. 2001. "The Organisation of Luo Conceptions of Intelligence: A Study of Implicit Theories in a Kenyan Village." *International Journal of Behavioral Development* 25 (4): 367–378. doi:10.1080/01650250042000348.

Guarini, Marcello. 2006. "Particularism and the Classification and Reclassification of Moral Cases." *IEEE Intelligent Systems* 21 (4): 22–28. doi:10.1109/MIS.2006.76.

Gul, Faruk, and Wolfgang Pesendorfer. 2006. "Random Expected Utility." *Econometrica* 74 (1): 121–146. doi:10.1111/j.1468-0262.2006.00651.x.

Hall, John Storrs. 2007a. *Beyond AI: Creating the Conscience of the Machine.* Amherst, NY: Prometheus Books.

———. 2007b. "Self-Improving AI: An Analysis." *Minds and Machines* 17 (3): 249–259. doi:10.1007/s11023-007-9065-3.

———. 2011. "Ethics for Self-Improving Machines." In Anderson and Anderson 2011b, 512–523.

Halpern, Diane F., Anna S. Beninger, and Carli A. Straight. 2011. "Sex Differences in Intelligence." In Sternberg and Kaufman 2011, 253–272.

Hanson, Robin. 2009. "Prefer Law to Values." *Overcoming Bias* (blog), October 10. http://www.overcomingbias.com/2009/10/prefer-law-to-values.html.

Hare, Richard Mervyn. 1952. *The Language of Morals.* Oxford: Clarendon.

———. 1982. "Ethical Theory and Utilitarianism." In *Utilitarianism and Beyond,* edited by Amartya Sen and Bernard Williams, 22–38. New York: Cambridge University Press. doi:10.1017/CBO9780511611964.003.

Harsanyi, John C. 1977. "Rule Utilitarianism and Decision Theory." *Erkenntnis* 11 (1): 25–53. doi:10.1007/BF00169843.

Hess, Stephane, and Andrew Daly, eds. 2010. *Choice Modelling: The State-of-the-art and the State-of-practice—Proceedings from the Inaugural International Choice Modelling Conference.* Bingley, UK: Emerald Group.

Hibbard, Bill. 2012. "Model-Based Utility Functions." *Journal of Artificial General Intelligence* 3 (1): 1–24. doi:10.2478/v10229-011-0013-5.

Honarvar, Ali Reza, and Nasser Ghasem-Aghaee. 2009. "An Artificial Neural Network Approach for Creating an Ethical Artificial Agent." In *2009 IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA),* 290–295. Piscataway, NJ: IEEE Press. doi:10.1109/CIRA.2009.5423190.

Hurka, Thomas. 1993. *Perfectionism.* Oxford Ethics Series. New York: Oxford University Press.

Hursthouse, Rosalind. 2012. "Virtue Ethics." In *The Stanford Encyclopedia of Philosophy,* Spring 2012, edited by Edward N. Zalta. Stanford University. http://plato.stanford.edu/archives/spr2012/entries/ethics-virtue/.

Idel, Moshe. 1990. *Golem: Jewish Magical and Mystical Traditions on the Artificial Anthropoid.* SUNY Series in Judaica. Albany: State University of New York Press.

Jackson, Frank. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis.* New York: Oxford University Press. doi:10.1093/0198250614.001.0001.

Jackson, Frank, and Michael Smith. 2006. "Absolutist Moral Theories and Uncertainty." *Journal of Philosophy* 103 (6): 267–283. http://www.jstor.org/stable/20619943.

Johansson, Petter, Lars Hall, Sverker Sikström, and Andreas Olsson. 2005. "Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task." *Science* 310 (5745): 116–119. doi:10.1126/science.1111709.

Johnson, Liz. 2009. "Are We Ready for Nanotechnology? How to Define Humanness In Public Policy." Paper prepared for the American Political Science Association (APSA) 2009 Annual Meeting, Toronto, ON, September 3–6. http://ssrn.com/abstract=1451429.

Johnson, Robert. 2010. "Kant's Moral Philosophy." In *The Stanford Encyclopedia of Philosophy,* Summer 2010, edited by Edward N. Zalta. Stanford University. http://plato.stanford.edu/archives/sum2010/entries/kant-moral/.

Joy, Bill. 2000. "Why the Future Doesn't Need Us." *Wired,* April. http://www.wired.com/wired/archive/8.04/joy.html.

Joyce, Richard. 2001. *The Evolution of Morality.* Cambridge Studies in Philosophy. New York: Cambridge University Press. doi:10.2277/0521808065.

Kaci, Souhila. 2011. *Working with Preferences: Less Is More.* Cognitive Technologies. Berlin: Springer. doi:10.1007/978-3-642-17280-9.

Kagan, Shelly. 1997. *Normative Ethics.* Dimensions of Philosophy. Boulder, CO: Westview.

Keeney, Ralph L., and Howard Raiffa. 1993. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs.* New York: Cambridge University Press. doi:10.2277/0521438837.

Kringelbach, Morten L., and Kent C. Berridge, eds. 2009. *Pleasures of the Brain.* Series in Affective Science. New York: Oxford University Press.

Laird, James D. 2007. *Feelings: The Perception of Self.* Series in Affective Science. New York: Oxford University Press. doi:10.1093/acprof:oso/9780195098891.001.0001.

Laurence, Stephen, and Eric Margolis. 2003. "Concepts and Conceptual Analysis." *Philosophy and Phenomenological Research* 67 (2): 253–282. doi:10.1111/j.1933-1592.2003.tb00290.x.

Legg, Shane. 2008. "Machine Super Intelligence." PhD diss., University of Lugano. http://www.vetta.org/documents/Machine_Super_Intelligence.pdf.

———. 2009. "On Universal Intelligence." *Vetta Project* (blog), May 8. http://www.vetta.org/2009/05/on-universal-intelligence/.

Legg, Shane, and Marcus Hutter. 2007. "A Collection of Definitions of Intelligence." In *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms—Proceedings of the AGI Workshop 2006,* edited by Ben Goertzel and Pei Wang, 17–24. Frontiers in Artificial Intelligence and Applications 157. Amsterdam: IOS.

Lewis, David. 1989. "Dispositional Theories of Value." *Proceedings of the Aristotelian Society, Supplementary Volumes* 63:113–137. http://www.jstor.org/stable/4106918.

Lim, Seung-Lark, John P. O'Doherty, and Antonio Rangel. 2011. "The Decision Value Computations in the vmPFC and Striatum Use a Relative Value Code That is Guided by Visual Attention." *Journal of Neuroscience* 31 (37): 13214–13223. doi:10.1523/JNEUROSCI.1246-11.2011.

Mackie, John Leslie. 1977. *Ethics: Inventing Right and Wrong.* New York: Penguin.

Mahoney, Matt. 2010. "A Model for Recursively Self Improving Programs v.3." Unpublished manuscript, December 17. Accessed March 27, 2012. http://mattmahoney.net/rsi.pdf.

McFadden, Daniel L. 2005. "Revealed Stochastic Preference: A Synthesis." *Economic Theory* 26 (2): 245–264. doi:10.1007/s00199-004-0495-3.

McLaren, Bruce M. 2006. "Computational Models of Ethical Reasoning: Challenges, Initial Steps, and Future Directions." *IEEE Intelligent Systems* 21 (4): 29–37. doi:10.1109/MIS.2006.67.

Minsky, Marvin. 1984. "Afterword to Vernor Vinge's novel, 'True Names.'" Unpublished manuscript, October 1. Accessed December 31, 2012. http://web.media.mit.edu/~minsky/papers/TrueNames.Afterword.html.

Moor, James H. 2006. "The Nature, Importance, and Difficulty of Machine Ethics." *IEEE Intelligent Systems* 21 (4): 18–21. doi:10.1109/MIS.2006.80.

Moore, George Edward. 1903. *Principia Ethica.* Cambridge: Cambridge University Press.

Moskowitz, Gordon B., Peizhong Li, and Elizabeth R. Kirk. 2004. "The Implicit Volition Model: On the Preconscious Regulation of Temporarily Adopted Goals." *Advances in Experimental Social Psychology* 36:317–413. doi:10.1016/S0065-2601(04)36006-5.

Muehlhauser, Luke. 2011. "So You Want to Save the World." Last revised March 2, 2012. http://lukeprog.com/SaveTheWorld.html.

———. 2012. "The Human's Hidden Utility Function (Maybe)." *Less Wrong* (blog), January 28. http://lesswrong.com/lw/9jh/the_humans_hidden_utility_function_maybe/.

Muehlhauser, Luke, and Anna Salamon. 2012. "Intelligence Explosion: Evidence and Import." In *Singularity Hypotheses: A Scientific and Philosophical Assessment,* edited by Amnon Eden, Johnny Søraker, James H. Moor, and Eric Steinhart. The Frontiers Collection. Berlin: Springer.

Neisser, Ulric. 1979. "The Concept of Intelligence." *Intelligence* 3 (3): 217–227. doi:10.1016/0160-2896(79)90018-7.

Nielsen, Thomas D., and Finn V. Jensen. 2004. "Learning a Decision Maker's Utility Function from (Possibly) Inconsistent Behavior." *Artificial Intelligence* 160 (1–2): 53–78. doi:10.1016/j.artint.2004.08.003.

Niu, Weihua, and Jillian Brass. 2011. "Intelligence in Worldwide Perspective." In Sternberg and Kaufman 2011, 623–645.

Nozick, Robert. 1974. *Anarchy, State, and Utopia.* New York: Basic Books.

Omohundro, Stephen M. 2008. "The Basic AI Drives." In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference,* edited by Pei Wang, Ben Goertzel, and Stan Franklin, 483–492. Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS.

Padoa-Schioppa, Camillo. 2011. "Neurobiology of Economic Choice: A Good-Based Model." *Annual Review of Neuroscience* 34:333–359. doi:10.1146/annurev-neuro-061010-113648.

Parfit, Derek. 1986. *Reasons and Persons.* New York: Oxford University Press. doi:10.1093/019824908X.001.0001.

———. 2011. *On What Matters.* 2 vols. The Berkeley Tanner Lectures. New York: Oxford University Press.

Pettit, Philip. 2003. "Akrasia, Collective and Individual." In *Weakness of Will and Practical Irrationality,* edited by Sarah Stroud and Christine Tappolet. New York: Oxford University Press. doi:10.1093/0199257361.003.0004.

Pettit, Philip, and Michael Smith. 2000. "Global Consequentialism." In *Morality, Rules, and Consequences: A Critical Reader,* edited by Brad Hooker, Elinor Mason, and Dale E. Miller, 121–133. Edinburgh, UK: Edinburgh University Press.

Posner, Richard A. 2004. *Catastrophe: Risk and Response.* New York: Oxford University Press.

Powers, Thomas M. 2006. "Prospects for a Kantian Machine." *IEEE Intelligent Systems* 21 (4): 46–51. doi:10.1109/MIS.2006.77.

Pratchett, Terry. 1996. *Feet of Clay: A Novel of Discworld.* Discworld Series. New York: HarperTorch.

Railton, Peter. 1986. "Facts and Values." *Philosophical Topics* 14 (2): 5–31.

———. 2003. *Facts, Values, and Norms: Essays Toward a Morality of Consequence.* Cambridge Studies in Philosophy. New York: Cambridge University Press. doi:10.1017/CBO9780511613982.

Rangel, Antonio, Colin Camerer, and P. Read Montague. 2008. "A Framework for Studying the Neurobiology of Value-Based Decision Making." *Nature Reviews Neuroscience* 9 (7): 545–556. doi:10.1038/nrn2357.

Rangel, Antonio, and Todd Hare. 2010. "Neural Computations Associated with Goal-Directed Choice." *Current Opinion in Neurobiology* 20 (2): 262–270. doi:10.1016/j.conb.2010.03.001.

Reynolds, Carson, and Alvaro Cassinelli, eds. 2009. *AP-CAP 2009: The Fifth Asia-Pacific Computing and Philosophy Conference, October 1st-2nd, University of Tokyo, Japan, Proceedings.* AP-CAP 2009. http://ia-cap.org/ap-cap09/proceedings.pdf.

Ring, Mark, and Laurent Orseau. 2011. "Delusion, Survival, and Intelligent Agents." In Schmidhuber, Thórisson, and Looks 2011, 11–20.

Russell, Stuart J., and Peter Norvig. 2010. *Artificial Intelligence: A Modern Approach.* 3rd ed. Upper Saddle River, NJ: Prentice-Hall.

Ruzgis, Patricia, and Elena L. Grigorenko. 1994. "Cultural Meaning Systems, Intelligence and Personality." In *Personality and Intelligence,* edited by Robert J. Sternberg and Patricia Ruzgis, 248–270. New York: Cambridge University Press. doi:10.2277/0521417902.

Rzepka, Rafal, and Kenji Araki. 2005. "What Statistics Could Do for Ethics? The Idea of Common Sense Processing Based Safety Valve." In Anderson, Anderson, and Armen 2005, 85–86.

Sandberg, Anders, and Nick Bostrom. 2008. *Whole Brain Emulation: A Roadmap.* Technical Report, 2008-3. Future of Humanity Institute, University of Oxford. http://www.fhi.ox.ac.uk/Reports/2008-3.pdf.

Schmidhuber, Jürgen. 2007. "Gödel Machines: Fully Self-Referential Optimal Universal Self-Improvers." In *Artificial General Intelligence,* edited by Ben Goertzel and Cassio Pennachin, 199–226. Cognitive Technologies. Berlin: Springer. doi:10.1007/978-3-540-68677-4_7.

Schmidhuber, Jürgen, Kristinn R. Thórisson, and Moshe Looks, eds. 2011. *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3–6, 2011. Proceedings.* Lecture Notes in Computer Science 6830. Berlin: Springer. doi:10.1007/978-3-642-22887-2.

Schnall, Simone, Jonathan Haidt, Gerald L. Clore, and Alexander H. Jordan. 2008. "Disgust as Embodied Moral Judgment." *Personality and Social Psychology Bulletin* 34 (8): 1096–1109. doi:10.1177/0146167208317771.

Schroeder, Timothy. 2004. *Three Faces of Desire.* Philosophy of Mind Series. New York: Oxford University Press. doi:10.1093/acprof:oso/9780195172379.001.0001.

Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (03): 417–424. doi:10.1017/S0140525X00005756.

Sen, Amartya. 1979. "Utilitarianism and Welfarism." *Journal of Philosophy* 76 (9): 463–489. doi:10.2307/2025934.

Shafer-Landau, Russ. 2003. *Moral Realism: A Defence.* New York: Oxford University Press.

Shope, Robert K. 1983. *The Analysis of Knowing: A Decade of Research.* Princeton, NJ: Princeton University Press.

Shulman, Carl, Henrik Jonsson, and Nick Tarleton. 2009a. "Machine Ethics and Superintelligence." In Reynolds and Cassinelli 2009, 95–97.

———. 2009b. "Which Consequentialism? Machine Ethics and Moral Divergence." In Reynolds and Cassinelli 2009, 23–25.

Simon, Dylan Alexander, and Nathaniel D. Daw. 2011. "Neural Correlates of Forward Planning in a Spatial Decision Task in Humans." *Journal of Neuroscience* 31 (14): 5526–5539. doi:10.1523/JNEUROSCI.4647-10.2011.

Single, Eric. 1995. "Defining Harm Reduction." *Drug and Alcohol Review,* no. 14 (3): 287–290. doi:10.1080/09595239500185371.

Slovic, Paul, Melissa L. Finucane, Ellen Peters, and Donald G. MacGregor. 2002. "The Affect Heuristic." In *Heuristics and Biases: The Psychology of Intuitive Judgment,* edited by Thomas Gilovich, Dale Griffin, and Daniel Kahneman, 397–420. New York: Cambridge University Press. doi:10.2277/0521796792.

Smart, R. N. 1958. "Negative Utilitarianism." *Mind,* n.s., 67 (268): 542–543. http://www.jstor.org/stable/2251207.

Smith, Kyle, Stephen V. Mahler, Susana Pecina, and Kent C. Berridge. 2009. "Hedonic Hotspots: Generating Sensory Pleasure in the Brain." In Kringelbach and Berridge 2009, 27–49.

Smith, Michael. 2009. "Desires, Values, Reasons, and the Dualism of Practical Reason." *Ratio* 22 (1): 98–125. doi:10.1111/j.1467-9329.2008.00420.x.

Sobel, David. 1994. "Full Information Accounts of Well-Being." *Ethics* 104 (4): 784–810. http://www.jstor.org/stable/2382218.

———. 1999. "Do the Desires of Rational Agents Converge?" *Analysis* 59 (263): 137–147. doi:10.1111/1467-8284.00160.

Stahl, Bernd Carsten. 2002. "Can a Computer Adhere to the Categorical Imperative? A Contemplation of the Limits of Transcendental Ethics in IT." In *Cognitive, Emotive and Ethical Aspects of Decision Making & Human Action,* edited by Iva Smit and George E. Lasker, 13–18. Vol. 1. Windsor, ON: International Institute for Advanced Studies in Systems Research / Cybernetics.

Sternberg, Robert J. 1985. "Implicit Theories of Intelligence, Creativity, and Wisdom." *Journal of Personality and Social Psychology* 49 (3): 607–627. doi:10.1037/0022-3514.49.3.607.

Sternberg, Robert J., Barbara E. Conway, Jerry L. Ketron, and Morty Bernstein. 1981. "People's Conceptions of Intelligence." *Journal of Personality and Social Psychology* 41 (1): 37–55. doi:10.1037/0022-3514.41.1.37.

Sternberg, Robert J., and Elena L. Grigorenko. 2006. "Cultural Intelligence and Successful Intelligence." *Group & Organization Management* 31 (1): 27–39. doi:10.1177/1059601105275255.

Sternberg, Robert J., and Scott Barry Kaufman, eds. 2011. *The Cambridge Handbook of Intelligence.* Cambridge Handbooks in Psychology. New York: Cambridge University Press.

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

Sverdlik, Steven. 1985. "Counterexamples in Ethics." *Metaphilosophy* 16 (2–3): 130–145. doi:10.1111/j.1467-9973.1985.tb00159.x.

Tännsjö, Torbjörn. 1998. *Hedonistic Utilitarianism.* Edinburgh, UK: Edinburgh University Press.

Tanyi, Attila. 2006. "An Essay on the Desire-Based Reasons Model." PhD diss., Central European University. http://web.ceu.hu/polsci/dissertations/Attila_Tanyi.pdf.

Tegmark, Max. 2007. "The Multiverse Hierarchy." In *Universe or Multiverse?,* edited by Bernard Carr, 99–126. New York: Cambridge University Press.

Thorndike, Edward L. 1911. *Animal Intelligence: Experimental Studies.* New York: Macmillan.

Tonkens, Ryan. 2009. "A Challenge for Machine Ethics." *Minds and Machines* 19 (3): 421–438. doi:10.1007/s11023-009-9159-1.

Tversky, Amos, and Daniel Kahneman. 1981. "The Framing of Decisions and the Psychology of Choice." *Science* 211 (4481): 453–458. doi:10.1126/science.7455683.

Vogelstein, Eric. 2010. "Moral Reasons and Moral Sentiments." PhD diss., University of Texas. doi:2152/ETD-UT-2010-05-1243.

Wallach, Wendell, and Colin Allen. 2009. *Moral Machines: Teaching Robots Right from Wrong.* New York: Oxford University Press. doi:10.1093/acprof:oso/9780195374049.001.0001.

Wallach, Wendell, Colin Allen, and Iva Smit. 2008. "Machine Morality: Bottom-Up and Top-Down Approaches for Modelling Human Moral Faculties." In "Ethics and Artificial Agents." Special issue, *AI & Society* 22 (4): 565–582. doi:10.1007/s00146-007-0099-0.

Weatherson, Brian. 2003. "What Good are Counterexamples?" *Philosophical Studies* 115 (1): 1–31. doi:10.1023/A:1024961917413.

Wilson, Timothy D. 2002. *Strangers to Ourselves: Discovering the Adaptive Unconscious.* Cambridge, MA: Belknap.

Yudkowsky, Eliezer. 2001. *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures.* The Singularity Institute, San Francisco, CA, June 15. http://intelligence.org/files/CFAI.pdf.

———. 2004. *Coherent Extrapolated Volition.* The Singularity Institute, San Francisco, CA, May. http://intelligence.org/files/CEV.pdf.

———. 2008. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks,* edited by Nick Bostrom and Milan M. Ćirković, 308–345. New York: Oxford University Press.

———. 2011. "Complex Value Systems in Friendly AI." In Schmidhuber, Thórisson, and Looks 2011, 388–393.

Zhong, Chen-Bo, Brendan Strejcek, and Niro Sivanathan. 2010. "A Clean Self Can Render Harsh Moral Judgment." *Journal of Experimental Social Psychology* 46 (5): 859–862. doi:10.1016/j.jesp.2010.04.003.

Zimmerman, David. 2003. "Why Richard Brandt Does Not Need Cognitive Psychotherapy, and Other Glad News about Idealized Preference Theories in Meta-Ethics." *Journal of Value Inquiry* 37 (3): 373–394. doi:10.1023/B:INQU.0000013348.62494.55.