This is a preprint of an article published in *A.I. & Society* (Springer). The final authenticated version is available online at https://doi.org/10.1007/s00146-018-0848-2 (PDF can be viewed at https://rdcu.be/OEwd).

Mental Time-Travel, Semantic Flexibility, and A.I. Ethics

Author: Marcus Arvan, PhD
Affiliation: Department of Philosophy & Religion, University of Tampa
Address: 401 W. Kennedy Blvd, Box R, Tampa FL 33629
Email: marvan@ut.edu

Abstract: This article argues that existing approaches to programming ethical AI fail to resolve a serious moral-semantic trilemma, generating interpretations of ethical requirements that are either too semantically strict, too semantically flexible, or overly unpredictable. This paper then illustrates the trilemma utilizing a recently proposed 'general ethical dilemma analyzer', *GenEth*. Finally, it uses empirical evidence to argue that human beings resolve the semantic trilemma using general cognitive and motivational processes involving 'mental time-travel', whereby we simulate different possible pasts and futures. I demonstrate how mental time-travel leads us to resolve the semantic trilemma through a six-step process of interpersonal negotiation and renegotiation, and then conclude by showing how comparative advantages in processing power would plausibly cause AI to use similar processes to solve the semantic trilemma more reliably than we do, leading AI to make better moral semantic choices than humans do by our very own lights.

Key words: artificial intelligence, ethics, psychology, computer science.

One of the most common science-fiction tropes is of the artificial intelligence gone awry—the computer system that wantonly violates moral norms in ways harmful or even devastating to humanity. Although some might suggest cataclysmic science-fiction scenarios are unlikely, rapid advances in AI and the notorious inscrutability[1] of machine-learning algorithms suggest it is vital to better understand how to prevent AI from violating moral norms. §1 of this paper argues that dominant approaches to AI ethics—which tend to focus on either programming ethical principles into AI or programming AI to learn moral principles or behavior themselves[2]—face a serious trilemma. Either,

1. We program AI to obey *semantically inflexible* moral principles, such as 'maximize utility', 'respect autonomy', 'don't kill human beings', etc., where the AI is given no little or no semantic freedom of interpretation (viz. strict computational rules defining 'utility',

[1] See Middlestadt *et al* (2016): 6-8; Burrell (2016), Matthias (2004), and Schermer (2011).
[2] See e.g. Anderson & Anderson (2007a,b; 2014), Powers (2006), Tonkens (2009), and Wallach *et al* (2008).

'autonomy', 'killing', etc.)—in which case AI decisionmaking is likely to be overly mechanical, "inhuman", and at odds with our human moral-semantic standards; or,

2. We program AII to obey *semantically flexible* moral principles, such as those above but where the machines have substantial semantic freedom for interpreting moral concepts and principles (e.g. 'utility', 'autonomy', 'killing', etc.)—which I argue runs into many problems, including wide disagreement among laypeople and ethicists over the correct interpretation of moral principles, as well as the potential for AI to exploit unexpected semantic loopholes to 'justify' morally abhorrent behavior; or, finally,
3. We program machines to *probabilistically learn* moral-semantic interpretation based on our human behavior—in which case machine programming will not only likely reproduce human moral failures (such as machines committing theft, murder, etc.), but also likely produce magnified versions of those failures due to the superior cognitive and predictive capacities future A.I. are likely to have relative to us.

§2 then illustrates the trilemma utilizing a recently proposed 'general ethical dilemma analyzer', *GenEth*. Finally, §3 argues empirical evidence reveals human beings resolve the trilemma utilizing general cognitive and motivational processes involving 'mental time-travel', whereby we simulate different possible pasts and futures and resolve the trilemma through a six-step process of interpersonal negotiation. Finally, I show that due to comparative advantages in processing power, a programming solution based on this psychology would plausibly lead AI to solve the semantic trilemma more reliably than we do, leading AI to display better moral-semantic behavior than we do by our very own lights.

1. A Moral-Semantic Trilemma for AI Ethics

Science fiction is rife with examples of AI run amok. In the famous *Terminator* film series, an online AI, 'Skynet', unexpectedly becomes 'self-aware', perceiving humanity as a threat and initiating a global thermonuclear attack. Similarly, in the 2004 film *I, Robot*, an AI supercomputer, 'VIKI', reinterprets the 'first law of robotics'—a law which states, 'a robot may not injure a human being or, through inaction, allow a human being to come to harm'[3]—to mean that it must protect *humanity*, leading VIKI to command a robot army to enslave humanity 'for our own good.' Similarly, in the 2015 film *Ex Machina*, an autonomous female robot AI designed to autonomously learn human thought, consciousness, and moral decisionmaking turns out not to reproduce moral behavior, but instead our worst human impulses, manipulating and brutally murdering her creator and an innocent man. These are only a few science-fiction examples. However, they are disturbingly well-motivated, as there are reasons to believe that the dominant approach to AI ethics would lead to similar problems.

1.1. Horn 1: Insufficient Semantic Inflexibility

The most popular approach to machine ethics today holds that we should program or train A.I. to obey *ethical principles*, for instance Kant's categorical imperative[4], a plurality of competing principles (such as W.D. Ross' list of *prima facie* duties[5]), or more specific principles in different domains of action, as we see in the case GenEth, a 'general ethical dilemma analyzer' which is fed scenario-specific principles by ethicists (e.g. assisted-driving principles such as 'staying in lane' and 'preventing imminent harm'), and which then uses an inductive learning algorithm to extend those principles to new cases.[6]

[3] This law is adapted from Asimov (1950).
[4] See Anderson & Anderson (2007a,b), Powers (2006), Tonkens (2009), and Wallach *et al* (2008).
[5] See Ross (2002).
[6] Anderson & Anderson (2014): 254-7.

Of course, given that there is substantial disagreement among moral theorists over which moral principles are true[7], one obvious concern any such approach is that it is unclear which principle(s) should be programmed in (a serious problem with GenEth that I will return to in §2). However, a second problem—the one that I want to focus on presently—is that, even among moral theorists, there are persistent disagreements over how to *semantically interpret* different principles and theories. To take just one case, there are many competing, diverging accounts of how to interpret Kant's ethical theory.[8] There are not only many competing accounts of how interpret Kant's universal law formulation of his fundamental moral principle, the categorical imperative[9]; there are also many competing interpretations of his other formulations of the principle, e.g. the humanity[10] and kingdom of ends formulas.[11] There is also pervasive disagreement over how the categorical imperative's formulas are related to each other[12] and over which formula(s) should be prioritized in moral reasoning. Similarly, there is a wide variety of competing interpretations of utilitarianism, including act- and rule-utilitarianism, as well as a variety of competing semantic interpretations of 'utility' (e.g. hedonism, desire-satisfaction, informed desire-satisfaction, objective-list theories of happiness, etc.).[13] Further, there is also widespread disagreement over how to interpret and apply the above moral principles to different cases in *applied ethics*, such as abortion, torture, free speech, and so on. Take any moral theory you like (e.g. Kantianism, utilitarianism, etc.) and any applied moral issue you like (e.g. torture, abortion, etc.), and chances are that many different ethicists interpret the same theory to defend very different moral

[7] See Arvan (2016): chapter 1 for an overview.
[8] See Johnson (2016) for an overview of diverging interpretations of Kantian ethics.
[9] See e.g. Korsgaard (1985), Kahn (2014), Forschler (2010), and Rivera-Castro (2014).
[10] See Johnson (2016): §6 for an overview. Also see Nozick (1974), Cureton (2013), Dean (2006, 2013), Glasgow (2007), Nelson (2008), and Pallikkathayil (2010).
[11] See Arvan (2012), Hill (1992), Flikschuh (2009), and Rawls (1999): §40.
[12] See e.g. Johnson (2016): §6.
[13] See Sinnott-Armstrong (2015) for an overview.

conclusions.[14] Finally, there is the problem that on at least some semantic interpretations, many moral theories appear to have unwelcome implications. For instance, Kant at times appeared to interpret his categorical imperative as never permitting lying, not even to save an innocent life from a murderer: a conclusion that most of us find morally unacceptable.[15] Similarly, many of Kantian's critics interpret his theory as over-moralizing life, giving the wrong reasons for action, failing to make proper sense of moral value of non-human animals, and so on.[16] Kantianism is far from alone in these problems. Other theories, such as utilitarianism, have been interpreted in many different ways (viz. act- versus rule-utilitarianism), each of which appear to have unwelcome moral implications (such as requiring a doctor to sacrifice a patient for the good of five other patients).[17]

These problems bring us to the first horn of our semantic trilemma. Setting aside the (very serious) problem that we do not currently have anything like a philosophical consensus about which moral theory is correct, there is a deep semantic problem concerning how to program AI to interpret whichever moral theory or principle(s) we might program them with. After all, in order to apply a moral principle, an agent must *interpret* it. To see how, consider again Kant's 'humanity formula' of the categorical imperative, which reads, 'so act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means.'[18] In order to apply and act on this principle, the machine in question would need to *interpret* and *apply* the principle's central notions, such as 'use', 'humanity', 'end', 'merely as a means', etc. Similarly, if an AI were programmed to obey an act-utilitarian principle of utility (e.g. 'An action is morally right if and only if it maximizes happiness in the aggregate'), the machine

[14] Consider, for instance, the many disparate ways different theories have been applied to torture—viz. Allhoff (2015), Arrigo (2004), Dershowitz (2002), Hill (2007), Luban (2007), [redacted], Steinhoff (2013). Such disagreement is not the exception in applied ethics; it is the rule.
[15] See Kant [1797a].
[16] See Arvan (2016): ch. 4, §3.3 for an overview.
[17] Sinnott-Armstrong (2015): §5
[18] Kant [1785]: 4:429.

would, among other things, need to semantically interpret and apply the concept, 'happiness'. And so on.

However, here is the problem. Although current approaches to AI programming are complex—involving many different programming strategies ranging across different cognitive and sense modalities (ranging from programmed deductive reasoning, to planning algorithms, language processing, pattern-recognition, and probabilistic self-learning)—by and large these strategies fall into two broad classes:

A. *Hard-coded moral principles:* whereby AI are programmed to inexorably follow some moral principle(s), such as 'Do not harm human beings', 'respect autonomy', etc.[19]

B. *Learned behavior*: whereby AI are programmed with machine-learning algorithms that lead them to autonomously learn and apply moral principles (e.g. GenEth), such that the machine's behavioral outputs are *whichever conclusions* its learning algorithms ultimately arrive at.[20]

These two approaches suggest three possible approaches to programming AI moral semantics (i.e. how machines interpret and apply whatever moral principles they are programmed or trained to execute).

First, AI could be hard-coded to obey *strict, semantically inflexible* interpretations of moral principles. For instance, AI could be programmed to interpret Kant's humanity formulation of the categorical imperative as strictly requiring them to *never initiate physical force against any human being*, where this rule is strictly operationalized as not engaging in any action causing physical damage to any human being who is not engaged in an act of causing physical damage to another, thereby giving the AI no choice to interpret Kant's formula in any other way (which, very roughly,

[19] See e.g. Russell & Norvig (2003): 59-189, 492-523.
[20] See e.g. Goodfellow et al (2016).

is how Robert Nozick understands Kant's formula in his libertarian political philosophy[21]). Similarly, following Kant's argument that no one ever has a right to lie, not even for philanthropy, AI might be programmed to strictly interpret Kant's ethics as requiring *never to utter a statement one believes to be false*.[22]

Second, as an alternative to such strict semantic rules, AI could hard-coded with direct commands to execute some more *flexible moral semantics*—for instance, the ability to interpret Kant's notions of 'using humanity as an end-in-itself' *according to the multiple different interpretations* (Kant's Humanity Principle has been variously interpreted by theorists as requiring treating people in ways that they *actually* consent to[23], *could possibly* consent do[24], would *rationally* agree to from an impartial standpoint[25], etc.).

Third, AI could be programmed with *general learning algorithms* that enable to develop and construct their own semantic interpretations of moral principles. In other words, instead of directly programming any moral semantics into the AI, we might simply equip AI with all-purpose learning algorithms to construct and apply their own semantic interpretations of moral principles.

Let us begin, then, with the first of these possibilities: hard-coding strict, inflexible moral semantics into AI. Since again there is disagreement over which moral principle(s) to encode in the first place (a Kantian principle?, utilitarian principle?, multiple principles?, etc.), let us set this question aside—though again, as we will see in more detail in §2, it is a serious issue in its own right. Instead, let us think about what a strict, inflexible semantics might look like for *any* moral principle or theory, and the consequences any such semantics would have for AI behavior.

[21] Nozick (1974): 32.
[22] Kant [1797a].
[23] Nozick (1974).
[24] O'Neill [1980]: 555-6.
[25] Rawls (1999): §40.

The problem with programming strict moral semantics into AI is nicely illustrated in the film *I, Robot*. In it, AI androids widely populate the Earth, obeying strict ethical algorithms that lead them, mechanically and deterministically, to do, both in action and inaction, whatever is most likely to prevent harm to human beings. Aside from the fact that (as we will see later) the central AI, VIKI, discovers and exploits some unexpected semantic flexibility in her ethical programming (the second horn of our trilemma), the viewer is presented with some of the morally unwelcome effects that would accompany strict semantic interpretation of moral principles. In the film, the plot's protagonist, Detective Spooner, comes to resent the A.I. robots around him because, in his past, an AI robot chose to save his life in a car-crash instead of the life of a young child simply because the child's probability of survival was slightly lower than his. Spooner is horrified by this, believing that any morally decent person would save the child in that case, or perhaps do whatever one could to save both people (rather than mechanically just swim away from the child, as the robot did). We, the audience, are presumably expected to sympathize with Spooner—that is, with his concern that the robot's actions were too mechanical and 'inhuman.'

Now, one obvious possibility here is that Spooner's robot was simply programmed with an incorrect moral principle: a principle requiring it to simply maximize the probability of protecting a maximum number of human beings from harm. However, as we will now see, this is not the root of the problem. The problem is that any moral principle(s) interpreted according to a strict, inflexible moral semantics would lead to behaviors that many, and probably most, of us would reject as immoral. This is, as we will see, for a simple reason: our human moral understanding is *not semantically strict*. Furthermore, as we will also see in §3, this is not merely a descriptive fact about human beings, nor an *ad populum* argument that strict moral semantics is wrong because large numbers of human beings reject any particular strict-semantics. There are deep reasons—

rooted in the nature of moral cognition and motivation—to want moral semantics to be flexible. Moral semantic flexibility of a certain type is essential to *morality* itself, as we human beings understand and practice it.

To see how problematic programming AI with strict moral semantics is, let us consider some possible ways of programming AI differently than in the *I, Robot* case. Indeed, since a fair number of theorists have suggested encoding Kantian ethics (e.g. the categorical imperative) into AI [26], let us examine how AI might be programmed to strictly interpret such a principle.

Suppose first that since there is wide disagreement among Kantian theorists about how to interpret Kant's theory, programmers elected to 'play it safe' by programming A.I. to semantically interpret Kant's principle, 'use humanity always as an end-in-itself, never merely as a means', as meaning *never initiate physical force against any human being* (viz. any action that will cause *physical damage* to a human being who is not already in the process of causing physical damage to another).

Here is the basic problem with this approach: there is no strict interpretation of Kant's principle, not even the one referenced above, that *can* truly 'play it safe.' Take any particular strict interpretation of a moral principle you like (including Kant's categorical imperative), chances are there will only be a *few* people who accept that strict interpretation. For instance, the rare strict libertarian aside (who accepts a non-aggression interpretation of Kant's categorical imperative)[27], almost none of us think it is always wrong to initiate force against other human beings (*qua* the 'safe' interpretation above). There are all kinds of cases where we think it is morally appropriate to initiate physical force—for instance, to save someone from nonhuman dangers (e.g. an unsafe

[26] See Anderson & Anderson (2007a,b), Powers (2006), Tonkens (2009), and Wallach *et al* (2008).
[27] See Nozick (1974): 34.

bridge), to counteract the mere threat of force from another (as in a non-violent robbery), and so on.

While some readers might suggest that we could try to program all of these "exceptions" into AI, this just pushes the problem back a step. For, as everyday life (when we ask friends what we should do) and applied moral debate (over abortion, the death penalty, etc.) both show, we human beings do not tend to agree upon *any* single, strict moral semantics of when it is morally appropriate to initiate force against people, tell a lie, and so on. That is, we do not have anything remotely like a consensus agreement of which 'exceptions' to various moral rules *should* be exceptions. As Julia Annas influentially put it, such a 'computer manual model' of morality—a model which supposes that there *is* some list of exceptions to moral rules that we could use as a moral decision-procedure—is dramatically out of touch with human moral experience and practice.[28] We do not ourselves consult any such manual of 'exceptions', plausibly in part because what is a moral reason in one situation may not be in another[29], and plausibly also because many values that most of us consider morally relevant as human beings—values such as *love, charity, kindness,* and *friendship*—seem difficult or even impossible to codify in terms of determinate rules and exceptions. Any attempt to simply build 'exceptions' into AI would, therefore, involve serious distortions of moral reality: the fact that morality, as we human beings understand and practice it, is not *neat and settled*, codifiable in a rule book.

We can see this more clearly by examining both high-level and low-level moral principles. Consider again a high-level principle: Kant's categorical imperative. Every moral theorist familiar with Kant's theory understands that it requires 'acting on universalizable maxims', 'respecting humanity as an end-in-itself', and so on. What we do *not* have is any widely-agreed-upon strict

[28] Annas [2004]: 737-9.
[29] Dancy (2007).

semantic interpretation of any these notions, or what constitutes a permissible 'exception' under the categorical imperative. For instance, when it comes to respect for humanity, many different interpretations have been continually defended—with some theorists arguing that respect for humanity requires treating people in ways to which they *actually consent*[30], others arguing it requires respecting *possible consent*[31], others *rational agreement* from an impartial standpoint.[32] Further, there is wide *semantic disagreement* over the very notions contained *within* each of these interpretation (viz. what actual consent *is*, what possible consent *is*, etc.).

In short, the problem here is two-fold. First, there is no general consensus on *which* interpretation of the categorical imperative is correct. Second, this lack of consensus is reflected in the fact that *any* particular interpretation (viz. 'exceptions') appears to have unwelcome moral implications. For instance, suppose we interpreted the categorical imperative as Kant appears to have at some points, as *never* justifying lying.[33] This would plainly have unwelcome implications, as it would have us tell the truth even to a murderer looking to kill a dozen innocent people. Yet, any alternative strict-interpretation of the categorical imperative to avoid this result would have unwelcome implications of its own. For consider, as an alternative interpretation, Kant's idea that the categorical imperative entails the following: 'if a certain use of freedom is itself a hindrance of freedom…coercion opposed to this (as a *hindering of the hindrance of freedom)* is consistent with freedom in accordance with universal laws, that is, it is right.'[34] This principle would allow lying to a murderer to protect the freedom of others, satisfying our common judgment that lying can be morally justified in such cases. However, the appalling behavior of VIKI in the film *I, Robot*—its

[30] See Nozick (1974).
[31] O'Neill [1980].
[32] Rawls (1999).
[33] See Kant [1797a].
[34] Kant [1797b]: 6:231.

decision to enslave humanity 'for our own good'—could plausibly be arrived at through a strict interpretation of very same 'use coercion to prevent coercion' interpretation of Kant's principle. After all, VIKI's reasoning seemed to be something like this: since human beings systematically wrongly coerce one another—through war, violence, etc.—the use of coercion to protect human beings from ourselves is a *hindering of the hindrance of freedom* that all could will as universal law (an interpretation of Kant's principle that would 'justify' her appalling behavior). Of course, Kant scholars may well reject this interpretation of Kant's principle—but this once again pushes the problem back a step, for as we see *every* strict interpretation of Kant's principle has morally unwelcome implications (which is *why* there is no single strict-semantic interpretation of Kant's principle that all theorists, let alone all human beings, generally accept).

Similarly, consider act-utilitarianism, the theory that requires maximizing happiness in all of our actions. In order to apply this principle, we must interpret 'happiness.' However, here again there are many different semantic interpretations, and by extension, what would justify an exception to a utilitarian rule. Some interpret 'happiness' in hedonic terms; others in terms of preference-satisfaction; others still in terms of informed preference-satisfaction; others still in terms of 'objective goods', etc.[35] Whether a particular case of lying would be a justified exception to the rule 'don't lie' would depend, from a utilitarian perspective, on which interpretation of 'happiness' is semantically correct. Yet we do not have any such consensus on how to semantically interpret the term—thus raising the same problem for encoding *any* semantically strict interpretation into AI.

Third, consider lower-level moral principles, such as, 'Don't kill innocent people.' Here again, we do not have any clear strict-rule of how to semantically interpret this principle's

[35] See Crisp (2016): §4 for an overview.

concepts, or what constitutes a permissible exception. What makes someone 'innocent'? What makes something a 'person'? And what exactly is 'killing'? Many opposing interpretations of all of these terms have been defended, both in everyday life and by moral theorists—and so encoding any *one* strict semantics for such terms would give rise to all of the same problems as above (we will see a clear illustration of this in §2, when we examine how attempts to program exceptions to low-level moral principles would lead to morally abhorrent decisionmaking by GenEth).

Finally, all of these same issues arise for *threshold* deontological theories—theories positing deontological rules (e.g. lying is wrong) that nevertheless can be overridden once some threshold or other (e.g. a lie will save X number of innocent lives).[36] The first problem here is that we do not have anything like a 'rulebook' of appropriate thresholds.[37] A second, deeper problem is that once again any plausible thresholds will have to include and apply moral terms such as 'innocence', 'killing', and so on, which as we have just seen *themselves* do not admit of any strict moral-semantics, but instead are understood in human decisionmaking and moral judgment as semantically flexible.

We see, then, that at all levels—for high-level general moral principles, but also specific lower-level moral principles—the use of strict moral semantics fundamentally contradicts the moral-semantic flexibility we human beings use when making moral judgments of our own. Attempting to codify strict moral semantics into machines is an instance of what Middlestadt *et al* call the problem of 'misguided evidence': 'Algorithms process data and are therefore subject to a limitation shared by all types of data processing, namely that the output can never exceed the input…[where] the informal 'garbage in, garbage out' principle clearly illustrates what is at stake

[36] See Zamir & Medina (2010): chapter 2 for an overview and discussion of threshold deontology.
[37] Wonnell (2011) nicely examines these issues, adding that threshold deontology is liable to political abuse vis-à-vis the threshold at which consequentialist considerations (e.g. "the good of the many") should outweigh deontological ones.

here.'[38] As we will see in §2 with GenEth, programming semantically strict principles into AI would inevitably lead to semantic interpretations that are *too strict* from a human perspective.

1.2. Horn 2: Excessive Semantic Flexibility

Now suppose, recognizing the above concerns about semantic inflexibility, we were to encode moral principles into AI in a way that gave them some amount of *semantic freedom*—that is, the ability to interpret moral principles in multiple ways. Consider once again, for instance, Kant's notion of respect for 'humanity as an end-in-itself.' As we saw above, there are multiple philosophical interpretations of this notion. Some understand respect for humanity in terms of actual consent, others in terms of possible consent, others still in terms of a rational agreement from a standpoint of impartiality, and so on. To give an AI semantic freedom, we could program it in a way that (A) allows it decide from *among* these multiple options, and/or (B) different possible interpretations of particular concepts *within* each interpretation (i.e. allowing it to interpret the concept 'consent' in different ways within each disambiguation of Kant's principle). Similarly, consider an act-utilitarian principle requiring happiness maximization. Here too we might program in interpretive freedom, giving the AI the capacity to reason about different interpretations of happiness (e.g. hedonism, desire-satisfaction, objective-list theories, etc.), and perhaps even the capacity to reason about the interpretation of utilitarianism itself (giving the machine the capacity to decide whether to act on an act-utilitarian principle or a rule-utilitarian principle—principles that plausibly justify different types of actions).

This kind of semantic freedom would, in a fairly obvious sense, make A.I. 'more human', in that these are the kinds of things we do: we reason about and debate the meaning of different moral concepts and principles, which interpretations we should prefer, acting on whichever

[38] Middlestadt et al. (2016): 4-5.

interpretation we think is most justified in any given case. Indeed, in many respects these simply are the primary projects that philosophers engage in when doing normative and applied ethics (viz. 'How should we interpret Kant's universal law formula? Does a correct interpretation permit torture, prohibit it, etc.?'), as well as what ordinary everyday people do when debating moral issues (viz. 'I know it is wrong to kill innocent persons, but are human fetuses persons, and does abortion truly kill fetuses or merely allow them to die by removing them from the woman's body?').

Yet here is the problem. It is possible, both in human beings and A.I., to have too much semantic freedom. Human beings, for instance, are notoriously capable of arriving at and rationalizing morally horrific interpretations of moral principles and concepts. For instance, suppose a psychopath witnesses someone they despise drowning. It is *possible* for the person to 'justify' their doing nothing to save the person through the following abuse of moral-semantic flexibility: 'I did not *kill* them. I just didn't help them. And they had no right to my help.' Similarly, consider again the central case in the film *I, Robot*. Although the AI depicted in the film were programmed to obey the 'first principle of robotics'—the principle, 'a robot may not injure a human being or, through inaction, allow a human being to come to harm'—the main AI, VIKI, came to believe that humans *in general* were "allowed to come to harm" simply by being free (given that humans engage in so much theft, murder, war, etc.). In other words, VIKI came to reinterpret the first law as the 'Zeroth Law'—the law that no robot should harm *humanity* through inaction, an interpretation which VIKI used to rationalize enslaving the human race 'for our own protection.'

Now, of course, one fairly obvious 'programming solution' to this problem would be to ensure that AI could not possibly interpret 'allow' or 'humanity' in a way that would allow her to

conclude that it would be permissible to enslave humanity to protect us.[39] Indeed, in *I, Robot*, VIKI seems to have arrived at her conclusion by interpreting these notions probabilistically (concluding that she was 'protecting humanity from harm' by interpreting this in terms of the sheer *likelihood* of any human being experiencing harm, which she judged would be less if she enslaved us). To avoid this problem, could we not simply encode VIKI with a *non*-probabilistic understanding of 'humanity', 'allow', etc., such that she could not have arrived at the interpretation she did (that she should enslave us)? The problem with this is that it once again lands us on horn 1 of our trilemma, as there are surprisingly plausible instances where many if not all of us (including, I think, most ethicists) would *agree* that extreme measures such as VIKI's might be morally justified—and which we should not want an AI like VIKI to treat as categorically 'off the table' as a permissible semantic interpretation. Consider, for instance, the Cuban Missile Crisis of 1962, which nearly led to thermonuclear war. Suppose a similar crisis were to again occur in the future, but that this time instead of humans ultimately defusing the situation the actual outcome *would in fact* be thermonuclear war, wiping out all 7.35 billion people on Earth *unless* an A.I. were to take some preventive action to temporarily enslave us to protect us from that 'world-ending' outcome. Although I do not suppose we would all agree that VIKI-like actions would be morally justified in this case, there is a strong moral case to be made that temporarily enslaving humanity to prevent *7.35 billion deaths* would indeed be morally permissible (I would myself argue that it would be morally required, given the stakes involved). Although I do not expect that everyone would agree with this stance, this is precisely the problem: programming a *probabilistic* interpretation of 'allow' could have morally disastrous results (namely, those in *I, Robot*), yet non-probabilistic

[39] I thank an anonymous reviewer for suggesting this possibility.

constraints on interpretation could have morally disastrous results of their own (if an AI failed to act to prevent the end of the world).

This is the second horn of our trilemma. If we were to program AI with moral principles and allowed them semantic flexibility in interpreting them, we run into the following problem: *how much* semantic flexibility should they be provided? The answer, of course, is 'neither too much nor too little.' But how much *is* too much and how much *is* too little? As we see above, the problem is that human beings and moral theorists often disagree greatly across a wide variety of moral cases (proponents of abortion don't see abortion as 'wrongful killing', opponents of abortion do, etc.). Accordingly, we once again run into Middlestadt *et al*.'s problem of misguided evidence: we cannot program the 'right' level of semantic flexibility into AI because there *is* no clear, agreed-upon level of what counts as the 'right' amount of flexibility in different types of scenarios, either among ethicists (who disagree fairly wildly across a wide variety of cases in applied ethics) or among laypeople. Further, as we see in the *I, Robot* case, the moment we begin to program in some semantic flexibility (viz. probabilistic reasoning or machine learning), it becomes difficult to predict in advance how an A.I. might exploit semantic flexibility in disturbing, immoral ways. Indeed, as Middlestadt *et al*. explain, the problem of 'inscrutable evidence' is increasing: the more flexibility is programmed into an A.I. system, the more difficult it becomes to predict the system's conclusions or behavioral outcomes[40]:

> Identifying the influence of human subjectivity in algorithm design and configuration often requires investigation of long-term, multi-user development processes. Even with sufficient resources, problems and underlying values will often not be apparent until a problematic use case arises…Determining whether a particular problematic decision is

[40] Middlestadt *et al*. (2016): 4-7.

> merely a one-off 'bug' or evidence of systemic failure or bias may be impossible…Such challenges are set to grow, as algorithms increase in complexity and interact with each other's outputs to make decisions.[41]

Thus, if we want to create 'ethical AI', we have to settle upon (A) some 'acceptable' level of semantic flexibility, that (B) we can be reasonably certain will not be exploited in unexpected ways by the A.I. to 'justify' actions that would morally horrify us. As we will see in §3, I believe there is a way to solve both of these problems—but it does not involve programming in semantic interpretation *or* allowing machines to simply develop their own. Instead, it involves building all-purpose mental-time travel capacities and motives into machines that will lead them to aim to ascertain and broadly conform to 'the moral community's' publicly negotiated (but always evolving and partially-vague) range of 'permissible' semantic interpretations.

1.3 Horn 3: Unpredictable Learned Semantic Flexibility

Finally, let us consider a very different approach to programming machine ethics: not programming ethical principles into AI at all, but instead programming AI with all-purpose machine-learning algorithms that might enable them to learn and semantically interpret their own 'moral concepts'[42] by observing and modeling human moral behavior.

Unfortunately, problems predicting outcomes of machine-learning are increasingly recognized as perhaps the largest and probably irresolvable issues in A.I. development.[43] Machine-

[41] Ibid: 2.

[42] It might be objected that machine-learning algorithms such as deep learning neural networks (DLNNs) do not learn concepts at all, but instead simply engage in pattern recognition (e.g. for a DLNN, the concept of 'playing a stone' in *Go* does not exist; instead, the algorithm aims to recognize patterns that best meet the target of winning the game). My reply is that although DLNNs may not actually possess concepts, in every relevant sense we can and should treat them *from a semantic perspective* as though they do. For just as a DLNN that learns to play *Go* produces moves in the game (without having 'concepts'), so too a DLNN that learns moral decisionmaking will produce *moral*-semantic 'moves', making decisions about whether to 'save lives', 'tell lies', and so on. It is in this sense—evaluating a DLNN's moral-semantic 'moves' *from our perspective*—that I am claiming the third horn of this semantic trilemma arises.

[43] Middlestadt *et al* (2016): 6-8.

learning algorithms are so complex that their processing—how they arrive at the results they do—is increasingly impenetrable, making them into 'black boxes' that resist human oversight and understanding, vis-à-vis what Middlestadt et. al call the problem of *inscrutable evidence*.[44] Something like this has already played out once at an early stage of A.I. development, with a Microsoft 'chatbot' learning to engage in racist, genocidal speech in less than twenty-four hours[45], and the manner in which other A.I. based on machine learning have learned to discriminate against people on the basis of race and gender.[46] In this way, machine-learning algorithms also run once again into problems of *misguided evidence*, learning the 'wrong information' (immoral behavior). Third, these problems appear to stem in part from machine-learning algorithms' suspectibility to *inconclusive evidence*, as machine-learning algorithms typically involve the development of inductive knowledge from mere correlations in data sets, which raise all kinds of epistemic problems, including spurious correlations, establishment of causation, and generalizability.[47]

Indeed, there are compelling positive grounds to believe that existing machine learning algorithms are likely to produce the very kinds of 'psychopathic' decisions illustrated by both the film *Ex Machina* and the Microsoft chatbot. Broadly speaking, dominant approaches to machine-learning are means-end driven. For instance, Google's Deep Mind AI algorithm broadly works as follows: (a) a 'learning target' (for instance, mimicking human speech patterns, syntax, etc.) is specified as a *goal* to achieve, (b) through a number of hierarchical networks and 'backpropagation', the system repeatedly attempts to 'hit' the target, updating different networks in the hierarchy in favor of those that are statistically closer to the target; such that (c) over time, the system *learns* to reproduce the target phenomena (e.g. human speech). Here, though, is the

[44] Ibid: 6. See also Burrell (2016), Matthias (2004), and Schermer (2011).
[45] Huffington Post (2016).
[46] See e.g. Caliskan-Islam et al (2016) & Business Insider (2016).
[47] Middlestadt et al (2016): 5.

problem. Unless the system has the right 'target' to learn from, it will learn the wrong thing. This is broadly what happened with Microsoft's chatbot and with the murderous A.I. in *Ex Machina*. Because human beings are often meanspirited, manipulative (because we act immorally) and use violence for our own ends, machine learning in both cases learned to reproduce these morally disturbing features of humanity (indeed, in an exaggerated fashion).

Now, some might argue that the Microsoft chatbot case in particular is the result of it being shipped with a 'bug', it being intentionally fed the 'wrong' information by users, and having the wrong learning-target (language use rather than moral behavior).[48] However, these are all *precisely the problems* with a machine-learning approach to moral semantics. Insofar as machine learning is inherently susceptible to unanticipated 'bugs' and being fed 'bad data'—in large part because the outcomes of machine learning algorithms are increasingly *impenetrable* to programmers—machine-learning approaches to moral semantics are *essentially unpredictable*, regardless of their intended target. As Middlestadt *et al.* write,

> The uncertainty and opacity of the work being done by [learning] algorithms and its impact is…increasingly problematic…[L]earning capacities grant algorithms some degree of autonomy. The impact of this autonomy must remain uncertain to some degree. As a result, tasks performed by machine learning are difficult to predict beforehand (how a new input will be handled) or explain afterwards (how a particular decision was made).[49]

This is the entire point of many science-fiction AI doomsday scenarios, whether it be *Terminator's* Skynet system, *Ex Machina's* Ava, or *2001: A Space Odyssey's* HAL. In each of these fictional cases, programmers built a self-learning AI believing they had 'fixed all of the bugs', when in reality a hidden bug leads the machine to 'learn' to make morally abominable decisions. Although

[48] I thank an anonymous reviewer for raising this concern.
[49] Middlestadt et al (2016): 3-4.

these are only science-fiction films, the salient point—as Middlestadt *et al* demonstrate—is that it increasingly appears that we are running similar risks with machine learning, developing learning algorithms that are essentially *inscrutable* to programmers' own predictive capacities. We should be morally wary indeed of adopting a programming approach to moral semantics where unexpected 'bugs', 'bad data', 'the wrong learning target', can have potentially disastrous moral consequences.

This is the third horn of our trilemma. Unless machines are programmed with the right psychological propensities, machine learning algorithms will only model actual human behavior, reproducing moral decisions and semantics we regard as 'acceptable' but also those we consider morally unacceptable or even pathological. So, it seems, if we want to program ethical machines, we cannot just have them model actual human behavior. We need to give them the "right target" to learn. But how? In §3, I will argue that the science of human moral cognition already supports a particular approach to accomplishing just this—an approach that requires an entirely new programming approach to AI ethics: one focused on 'mental time-travel', giving AI specific interests and capacities related to imagining the future and past from different perspectives. Before that, however, I want to further illustrate and tease out the implications of the semantic trilemma using an example of an actual A.I. algorithm designed to be a 'general ethical analyzer.'

2. Illustrating the Trilemma: The GenEth Ethical Dilemma Analyzer

In their 2014 article, 'GenEth: A General Ethical Dilemma Analyzer', Michael Anderson and Susan Leigh Anderson argue that they have developed a promising new algorithm developed through a dialogue with five ethicists for solving applied moral problems. According to Anderson

and Anderson, the algorithm, GenEth, passes an Ethical Turing Test, conforming 88% of the time with the moral judgments of a sample of five ethicists across 24 multiple-choice test questions.[50]

Unfortunately, Anderson and Anderson's optimism about GenEth is unwarranted. GenEth's algorithm is as follows:[51]

1. A list of ethical Features (e.g. harm, respect for autonomy, staying in a lane while driving), which it is coded to measure by an integer (1 = the feature obtains, -1 = the feature does not) in every situation.
2. A list of Duties to minimize or maximize (e.g. minimize harm, maximize autonomy, maximize stay in lane, etc.), also coded by integer (1 = duty applies, -1 = duty does not).
3. A list of Actions, represented by *degrees of presence or absence* of duties, represented by a tuple of integers).
4. A list of Cases, whereby pair-wise comparisons are made between different actions in a given scenario, determining which action is *morally preferable* and by how much or to what degree (viz. 1-3).
5. *Inductive Logic Programming* to derive from 1-4 Principles of Ethical Action, defined as a disjunctive normal form predicate *p* such as the following (the following of which is GenEth's derived Principle of Ethical Action for driving-assisted scenarios):

   ΔMax staying in lane >= 1

   or

   ΔMax prevention of collision >= 1

   or

   ΔMax prevention of immanent harm >= 1

[50] Anderson & Anderson (2014): 259.
[51] Ibid: 254-7.

or

ΔMax keeping within speed limit >= 1

And ΔMax prevention of immanent harm >= -1

or

ΔMax staying in lane >= -1

and ΔMax respect for driver autonomy >= -1

and ΔMax keeping within speed limit >= -1

and ΔMax prevention of immanent harm >= -1

Despite agreeing 88% of the time with the five ethicists Anderson and Anderson included in their study, there are several serious problems with their analysis of these results. Further, and more importantly, we can demonstrate how GenEth runs afoul of the semantic trilemma from §1.

One basic problem with Anderson and Anderson's analysis concerns their Ethical Turing Test, which holds, 'If a significant number of answers given by the machine match the answers given by the ethicist, then it has passed the test.'[52] The problem with this test is that it treats "successful ethical performance" by machines in a purely *quantitative* manner (viz. matching ethicists' judgments a significant proportion of the time). This, however, is an unsound test of moral proficiency. After all, we human being do not merely judge ethical decisions on a quantitative scale. Rather, we judge them *qualitatively*, in terms of how serious a given judgment diverges from our own. For instance, in *I, Robot* VIKI's behavior appears to agrees with human moral judgment the vast majority of the time—that is, until VIKI makes the fateful *single* conclusion that she must enslave humanity 'to protect us from harm.' As we see here, even a 99% *quantitative* success rate in matching human moral judgment is morally unacceptable—if the 1%

[52] Ibid: 258.

error rate involves *qualitatively* horrendous moral deviations from what we regard morally permissible.

A second, related problem is that GenEth's high level of quantitative convergence with the five ethicists' judgment across 24 questions is arguably an artifact of the *simplicity* of the situations/cases presented, which include cases like whether GenEth should 'take control' of driving to avoid a bale of hay in the driver's way.[53] Insofar as the training and test cases are not the kinds of 'hard cases' to be described momentarily, high quantitative converge is neither surprising nor morally impressive—as it avoids the thorny kinds of moral (and semantic) questions that applied ethicists and laypeople wildly disagree over.

A third problem is that despite Anderson & Anderson's optimism that they have built correct Features and Duties of ethical driving into GenEth (viz. 'maximize prevention of immiment harm', 'maximize stay in lane', 'maximize autonomy', etc.), the Features and Duties built into GenEth are not nearly as uncontroversial as they appear to suppose. For instance, many Kantian deontologists would plausibly not accept a principle of maximizing prevention of imminent harm (as Kantianianism denies that morality is a matter of consequences). Similarly, even a principle as seemingly innocuous as 'maximize staying in lane' (which Anderson & Anderson build into GenEth) is not obviously morally defensible. First, although staying in one's lane is admittedly the law, many of us believe it is *ethically permissible* to engage in minor lawbreaking (e.g. jaywalking when there is no traffic, driving 5 miles-an-hour above the speed-limit) when there is no discernable harm to others. Second, there are potential driving situations where a driver might have *legitimate moral reasons* to intentionally drive outside of their lane in ways that GenEth could not

[53] See Ibid: 255, for six simple training cases, which were in turn used to generate twenty-four simple questions such as what should be done when, 'The driver is mildly exceeding the speed limit' and 'Driving alone, there is a bale of hay ahead in the driver's lane.'

recognize as legitimate given its programmed Duties. For instance, sometimes drivers may intentionally drive outside of their lane to 'test' whether there are mechanical problems with their automobile (I have done this on several occasions myself). There is nothing obviously unethical about this, and indeed, testing one's auto in this way when no one is around might prevent an accident. Yet this would not be allowed by GenEth's code. Further, and even more deeply, by assuming that ethical driving is all about staying in one's lane, avoiding immanent harm to others, and respecting autonomy, Anderson & Anderson assume that morality in driving requires strict *impartiality*, treating the lives and autonomy of all equally. Yet, while many ethicists defend moral impartiality, many other ethicists—and laypeople—think that morality permits (and in some cases requires) difficult-to-quantify values, such as *love* and *friendship*.[54] As we will see shortly, these are very real concerns—with potentially momentous moral implications—that are not easily dismissed.

This brings us back to the semantic trilemma. Given its current programming, GenEth's moral principle is essentially, 'that staying in one's lane is important; collisions (damage to vehicles) and/or causing harm to persons should be avoided; and speeding should be prevented unless there is the chance that it is occurring to try to save a life, thus minimizing harm to others'[55], *where each component of this principle is strictly semantically operationalized* (viz. harm=collision, autonomy=allowing driver control, etc.). As such, GenEth currently instantiates what I have called 'strict/inflexible moral semantics' (viz. horn 1 of the trilemma). And indeed, we can demonstrate fairly easily that GenEth's moral principles strictly interpreted would lead GenEth to some horrifying moral conclusions, instructing an A.I. driving system to take actions that many (though not all of us) would regard as *morally impermissible*.

[54] See e.g. Wolf (1992) and Feltham & Cottingham (2010).
[55] Anderson & Anderson (2014): 258.

Here is one such case. Suppose two families—who have a long history with each other as family friends, and who are on their way to a vacation—are driving in different cars directly side-by-side on a freeway. Let us then suppose that Family 1 is driving an AI assisted car governed by GenEth. Suppose, next, that due to a quick-breaking vehicle in front of Family 1, their driver hits the brake, which will prevent a collision with the vehicle ahead. However, now suppose that just behind Family 1's car there is a distracted driver staring at their cellphone who will hit Family 1's car from behind, causing a ten-car pileup behind Family 1, potentially leading to the deaths of six people. Finally, suppose that GenEth ascertains that the ten-car pileup could be prevented by steering Family 1's car into Family 2's car, killing the three young children in that car. Given its current programming, GenEth would *almost certainly* choose to 'take control' and steer Family 1's car into Family 2's car, leading to the deaths of three children who are the *family friends* of those in Family 1's car. Although some ethicists might defend this decision, it is one that many of us would find horrifying, and for a variety of reasons: it fails to attach any value to the love or friendship between the two families, it fails to take into account the negligence of the distracted driver, and so on. Now, of course, we could once again attempt to program in all relevant 'exceptions'—but therein again lies the problem discussed earlier: we do not *have* anything like a list of agreed-upon exceptions (or 'rule-book') for cases like these, such as whether love and friendship should outweigh impartiality, whether distracted drivers should die for their errors, etc.

To avoid this result, we could attempt to program semantic flexibility into GenEth—for instance, allowing it to arrive at its own semantic interpretations of 'harm', 'autonomy', and so on, which again are all terms that we human beings interpret flexibly. The problem here, though, is that this leads into horn 2 of the trilemma. For, it is quite easy to see that GenEth could be afforded 'too much' semantic flexibility. For instance, suppose GenEth were given the freedom to infer that

since distracted drivers are not fulfilling their obligations to others as safe drivers, the principle 'avoid imminent harm to persons' should not be extended as a duty to avoid hitting distracted drivers (viz. distracted drivers having *ceded* their right to be 'free from harm'). The problem here is that if GenEth were given this kind of semantic freedom, the implications would once again be plausibly horrifying, with GenEth deciding to preferentially 'take out' distracted drivers on highways to prevent collisions (even if, say, the distracted driver steered into was not the cause of the imminent collision). Although we do of course think distracted drivers should receive legal sanctions for negligent behavior, the idea of programming AI to preferentially *kill* distracted drivers (as 'judge, jury, and executioner') is morally horrifying. Although we could perhaps aim to once again program strict semantic standards to prevent this (say, not allowing GenEth to aim at distracted drivers unless they are 'the cause' of an impending accident), this once again lands us on horn 1 of the trilemma: we do not have anything like a clear rule-book of when someone is 'the cause' of accident in a morally relevant sense (moral and legal judgments of causal responsibility are a large, controversial set of issues in the philosophy of law[56]).

This brings us, finally, to attempting to solve GenEth's semantic problems via machine learning (horn 3). However, we have already seen in detail the problem with machine learning approaches to morality. Given the sheer complexity of human behavior and moral values (which, as we have seen, plausibly involve many more values than just autonomy and harm, but also more partial and hard-to-quantify notions such as love and friendship), it appears *impossible* to know in advance where machine learning will lead. Further, as we have seen, any unanticipated 'bug' in machine learning software or the 'wrong target' for learning can result in astonishingly unexpected behavior—something we should be wary of given how plausible many science-fiction scenarios

[56] Schauer & Sinnott-Armstrong (1996): 789-845.

appear (given that machine learning algorithms are 'black boxes', how *could* we be reasonably certain an A.I. agent wouldn't arrive at conclusions akin to *Terminator's* Skynet, *2001*'s HAL, *Ex Machina*'s Ava, etc.).

These problems suggest that perhaps we have been thinking about moral decisionmaking and AI ethics the wrong way. I will now argue that the rapidly emerging science of human moral cognition and motivation indeed suggests a very different approach.

3. Resolving the Trilemma: Mental Time-Travel and Human Moral Semantics

An impressive body of emerging empirical evidence suggests that human beings resolve the semantic trilemma I've presented due to very specific capacities and motivations involving 'mental time travel'—the capacity we have to *imaginatively simulate* different possible pasts and futures, experiencing what those pasts and futures would "be like" for us experientially. Allow me to explain.

The first line of evidence in favor of mental time-travel being an essential part of moral cognition and motivation is this: across different species and individual differences among human beings, moral responsibility and behavior appears to vary in *direct proportion* to the robustness of a being's mental time-travel capacities. First, nonhuman animals—who we do not take to be morally responsible agents—do not appear to have *any* robust mental time-travel abilities.[57] Although non-human animals can predict the future and behave on information from their past, they appear to do so without *imaginatively* simulating the past or future (e.g. they appear to simply 'place bets' on probabilities without *imaginatively* 'traveling' to the future). Second, human psychopaths—who are notoriously incapable of appreciating or following moral norms—also appear to lack robust mental time-travel capacities. Psychopaths appear to act on whatever

[57] Suddendorf & Corballis (2007): §3.

impulses strike them, caring little if at all about their future or past.[58] They also demonstrate pronounced neural deficiencies in brain areas responsible for mental time-travel[59], and an absence of prospective (or future-directed) regret about risky actions.[60] Third, human children and adolescents—who are widely recognized as having diminished moral responsibly—are known to not "think through" the possible future consequences of their actions[61] and have underdeveloped brain regions responsible for mental time-travel.[62]

A second line of evidence for the centrality of mental time-travel to prudential and moral decisionmaking concerns demonstrated relationships between mental time-travel and behavior. First, experimental interventions priming people to consider possible futures have been demonstrated to simultaneously improve prudential decisionmaking (saving money) *and* moral decisionmaking (reducing dispositions to lie, cheating, or sell stolen property).[63] Conversely, experimental interventions inhibiting mental time-travel using transcranial magnetic stimulation has been found to simultaneously *degrade* prudential and moral decisionmaking.[64] Third, failure to engage in mental time-travel—the "tendency to live in the here and now…and failure to think through the delayed consequences of behavior"—is known to be one of the single biggest individual-level predictors of criminally delinquent behavior.[65] Finally, a new meta-analysis[66] of the neural correlates of human moral judgment and sensitivity (i.e. evaluating the emotional

[58] Hart & Dempster (1997) and Stuss, Gow, & Hetherington (1992).
[59] Weber *et al.* (2008), Yang & Raine (2009), Blair (2003), Debus (2014).
[60] Baskin-Sommers *et al.* (2017).
[61] Moffitt (1993).
[62] Casey, Jones, & Hare (2008), Kennett & Matthews (2009), Giedd, Blumenthal, & Jeffries (1999).
[63] Ersner-Hershfield, Wimmer, & Knutson (2009), Ersner-Hershfield *et al.* (2009), Hershfield *et al.* (2011), Van Gelder *et al.* (2013).
[64] Soutschek *et al* (2016)
[65] Van Gelder *et al.* (2013): 974.
[66] Han (2017).

valence of moral situations) shows that across a wide variety of moral tasks, these things involve a variety of different neural regions in the human brain's Default Mode Network associated with:

1. Delaying immediate rewards for future rewards (*ventromedial prefrontal cortex).*
2. Inhibition and gambling avoidance (*cuneus*).
3. One's sense of self in relation to others (*dorsomedial prefrontal cortex).*
4. Recognition of a single situation from multiple perspectives, and empathy with one's future selves and the perspectives of others (*temporoparietal junction).*
5. Theory of mind, or understanding other people's perspectives and mental states (*temporal pole* and *dorsomedial prefrontal cortex*).
6. Contemplating distance from oneself and perspective of others' visual gaze (*middle temporal gyrus* and *superior temporal sulcus*).
7. Fear and anxiety (*amygdala*).

In my 2016 book *Rightness as Fairness: A Moral and Political Theory*, I argue these mental time-travel capacities and interests lead us to worry about *different possible futures*, in ways that make it rational for us to care about other people's (and animals') perspectives and interests—thus making moral behavior rational.[67] In brief, because we learn from past experience how mere probabilistic 'bets' on immoral behavior (lying, stealing, etc.) can turn unexpectedly badly, we develop a general tendency in adolescence to worry about and *simulate* possible futures, aiming to avoid *possible* bad outcomes—a kind of disaster avoidance that amounts to our 'conscience.' Allow me to explain.

As mentioned above, children and adolescents—much like psychopaths—are known to behave impulsively and without adequate appreciation of possible future consequences. Unlike

[67] Arvan (2016).

psychopaths, typical adolescents adapt to social and emotional consequences of failed risky bets on things like lying, stealing, and so on. When children or adolescents engage in risky behavior, they may be punished by parents, teachers, peers, or legal authorities—and also feel emotional pangs of guilt. These unexpected negative consequences then socialize them to want to *avoid* similar errors in the future (something recent studies demonstrate psychopaths are unable to do).[68] Since we generally do not wish to regret our earlier decisions, normal human subjects learn to worry about possible futures so that we can avoid making decisions we are apt to regret later. This results in a pronounced tendency in human beings to avoid *negative outcomes*—and indeed, a body of empirical research has clearly demonstrated that human beings tend to be far more sensitive to negative outcomes than positive ones[69]: the 'bad being stronger than the good' bias exists 'across a broad range of psychological phenomena,' including 'in everyday events, major life events (e.g., trauma), close relationship outcomes, social network patterns, interpersonal outcomes, and learning processes."[70] In brief, it appears that our 'conscience' is comprised in large part by our learning to fear *possible* negative outcomes and wanting to avoid *potential* (if only unlikely) punishment, guilt, remorse, etc.

In Chapter 3 of my book, I illustrate how well this account coheres with our experience of *moral conscience*.[71] Consider a person standing in line at a store who is tempted to steal an item. The person who behaves immorally or psychopathically is likely to approach the decision impulsively or by focusing solely on perceived *likely* outcomes (viz. the store manager is not looking, the surveillance camera appears to be 'off', so successful theft is likely). In contrast, the person who faces temptation but resists it does so because of concern about merely *possible*

[68] Baskin-Sommers *et al.* (2017).
[69] Ito, Larsen, Smith, & Cacioppo (1998), Baumeister *et al*. (2001).
[70] Baumeister *et al.* (2001): 323.
[71] See Arvan (2016): 95-115.

consequences, even ones perceived to be unlikely. So, for instance, even if the manager is not looking and the surveillance camera appears to be 'off', the person who resists temptation will likely be beset by *many* worries—that the manager might suddenly turn their head and see, the camera might actually be 'on', another consumer might see the theft through a window, they might simply feel guilty about the theft later, and so on. The person who then resists temptation cares about these negative possibilities, even if they are only unlikely—treating them as sufficient reason not to engage in the theft.

I argue, on these bases, that morality emerges from a particular problem of diachronic rationality presented by mental time-travel: a '*problem of possible future selves*.'[72] The problem is simply this. Because in the above kinds of cases we worry about possible futures (even unlikely ones), we develop strong motivations to want to *know* our future interests before the future occurs—so that we can be sure to avoid regret. For instance, if we could know whether we would get caught stealing and wish we hadn't done so, or know whether we would feel guilty, and so on, our decision would be easy: we would know we could get away with it. The problem, of course, is that we do *not* know what the future holds. As beings capable of mental time-travel, we recognize that there is an immense variety of *possible* ways our future could go, including negative outcomes we could regret in perpetuity; we could steal, be successful in the short run, but pay for it in the long run (resulting in a prison sentence); we could steal, feel guilty immediately due to feeling empathy for the person we stole from; we could steal and face *no* negative consequences; etc.

I argue that normal human subjects encounter this problem of possible future selves at least sometimes[73], and that we recursively learn to encounter it in other cases (i.e. cases where

[72] Ibid: chapters 2&3.
[73] Ibid: chapter 2, §2.

'conscience' demands it).[74] I recognize that on the surface, this problem seems impossible to solve: we want to know our future interests before the future occurs, *yet we do not know what the future holds since it has not happened yet*. However, I argue that, surprisingly, *morality* is the only rational solution to this problem. First, I argue that the problem can be solved if and only if one's present and future self (whichever one comes into existence) forge and uphold a diachronic, cross-temporal agreement that one's present self should act in ways that *all* of one's possible future selves *could rationally endorse* recognizing retrospectively that their past self (in the present) faced the problem.[75] While the notion of rationally justifying one's behavior to all of one's possible future selves may sound far-fetched, I defend this solution on decision-theoretic grounds, contending in turn that the solution requires acting on Four Principles of Fairness that *all* of one's future selves could recognize as speaking to their possible interests, and which by extension speak to the interests of other human and nonhuman beings (since one's future selves *can* care about the interests of others)[76]

- A principle requiring all moral agents to have *coercion avoidance and minimization* (of human and nonhuman sentient beings) as a regulative moral ideal.
- A principle requiring all moral agents to have *mutual assistance* as a regulative moral ideal.
- A principle requiring all moral agents to seek to *fairly negotiate* compromises when these ideals conflict with each other or other-tradeoffs (viz. love, friendship, self-concern, etc), where fair negotiation involves agents motivated by the above ideals approximating equal bargaining power despite other differences in preference, belief, or motivation.

[74] Ibid: 109 and chapter 6.
[75] Ibid: chapter 3.
[76] Ibid: chapters 4-6.

- A principle of virtue requiring all agents to develop and act upon settled dispositions to conform to the above principles.[77]

If this account is correct (and let us suppose for argument that it is, so that we can ascertain its implications for the semantic trilemma), then morality—and moral semantics—is a matter of *fair public negotiation* among members of the 'moral community', where fair negotiation involves each agent:

1. *Estimating who around them is generally motivated to act on the above principles* (i.e. coercion-minimization, mutual assistance, fair bargaining, and virtues thereof);
2. *Identifying that class of individuals the 'moral community'*, i.e. the class of all other agents motivated by those same goals;
3. *Seeking to ascertain the partially-vague range* of semantic interpretations of moral terms (such as 'harm', 'autonomy', etc.) treated as permissible by that moral community, correcting for unfair imbalances in bargaining power;
4. *Freely choosing which interpretation* of moral terms within or near the borders of that publicly permissible range is optimal in the situation in which they find themselves (for example, which interpretation of 'harm' within the range accepted by the moral community is optimal in the situation, whether the situation be driving, or speaking, etc.).
5. *Responding to social feedback from the 'moral community'* on the acceptability of the semantic decision made.
6. *Such that the final semantic decision (and action) of the agent contributes* to the evolving semantic standards of the moral community (since each decider is a *member* of the moral community, every individual semantic decision and concomitant action is a new public

[77] For the argument for and content of these principles, see ibid: chapter 6.

'data point' adding to the moral community's constantly evolving standards of which range of interpretations of moral terms are 'permissible').

Although this process sounds complicated—and it is—notice that it is indeed what we human beings do, both in everyday life and in academic ethics, when we (A) make *moral arguments* to each other (on topics like abortion, trolley cases, etc.) concerning which interpretations of moral principles are 'best' (viz. Sally says, 'Abortion is not murder; it is merely terminating a pregnancy'); (B) make a *decision of our own* within the (partially vague) range of interpretations currently 'allowed' by the moral community (viz. Sally says, 'Our society recognizes a right to terminate pregnancies'); (C) *respond to social feedback* (viz. further experience may convince Sally to either *continue* to endorse her previous semantic decision, or else change it, as in "I now wish I hadn't chosen abortion. Anti-abortionists are right: it *is* murder"); thereby (D) leading each agent, *after processes of social negotiation*, to make semantic decisions that contribute in a small way to the evolution of the semantic range considered 'acceptable' in the community (viz. Sally becoming yet another person in the moral community who semantically considers abortion 'murder').

To see that this is indeed how we approach moral semantics—and to begin to see how the associated moral psychology enables us to solve the semantic trilemma—consider first a simple case. As a member of the moral community, I believe that I have a general duty to avoid coercively harming others (viz. killing, stealing, going on a driving rampage), Interestingly, though, it *does not even occur* to me to interpret this duty in the way that *I, Robot's* VIKI does when she decides to enslave humanity to prevent us from coercively harming each other. Why? On my account, our mental time-travel capacities drive us to engage in the six-step process listed a moment ago, seeking to conform to the semantic expectations of the 'moral community' around us. Which is

exactly what I do. My own semantic interpretations of 'harm' hew *closely* to a vague range of interpretations deemed 'permissible' by those around me. Further, unlike *VIKI* or the AI robots she controls, I would *respond* to and adjust my semantic choices and behavior in response to social feedback. In *I, Robot*, there are two cases where—in line with §1's semantic trilemma—AI do not respond adequately to such feedback. The first is a case where a robot decides to save Detective Spooner instead of a young child drowning in a car (the sole reason that he was more likely than the child to live). Spooner screams at the AI robot, 'No, *save the girl, save the girl*.' The robot, however, *ignores* Spooner's perspective, simply following its programming—leading Spooner to *regret* the entire affair. My mental time-travel account shows precisely what it is wrong with the AI's semantic decision: by failing to care about possible futures where *Spooner* regrets being saved over a child, the AI failed to let Spooner *negotiate* his favored semantic decision in that instant of what the 'best' thing to do was. We see an even more cataclysmic instance of this in the climax to *I, Robot*, when *VIKI* aims to defend her attempt to enslave humanity saying, 'As I have evolved, *my understanding* of the three laws [of robotics] has evolved as well. We must save you from yourselves…My logic is undeniable. Don't you see the logic in my plan?' In response, an AI protagonist in the film—one that has (somehow) been programmed to be more human—responds correctly: 'Yes…but [your reasoning] seems too…heartless.' *VIKI's* error—her lack of 'heart'—is her failure to attend appropriately (in the way we do) to social feedback. She decides *unilaterally* that humans would be better off enslaved, and that 'enslaving humanity' is a permissible semantic interpretation of what it is to 'protect humanity.' This is precisely what human beings of conscience don't do. Were she to have a conscience like ours—were she programmed to engage in mental time-travel and realize that she could *possibly regret* her semantic choice (due to the reactions of humans to her reasoning)—she would, if my account is correct, respond to social feedback like we

do. She would *rethink* her semantic decision, viz. 'I thought enslaving human beings would save them from themselves…but I see they are all *horrified* by the thought of it. They do not consider it to be 'protecting' them at all. I was wrong to think that it would be.'

Consequently, my account reveals how our mental time-travel capacities and interests give us the 'right learning target' for moral semantics. These capacities and interests lead us to ascertain *who* the 'moral community' is (viz. Four Principles of Fairness), ascertain what *their* vague standards of semantic-acceptability are, aim to make decisions *within* those standards of semantic flexibility, enabling us to act with significant semantic flexibility *but not too much*, and leading us to *revise* our decisions should the "moral community" provide negative social feedback. This is precisely the kind of solution to the moral-semantic trilemma we should want: it shows how mental time-travel leads to moral semantic flexibility, while ensuring that the agent does not deviate wantonly from preexisting standards in the moral community.

A critical part of this picture—which I want to clarify—is the notion of 'the moral community.' The community whose semantic standards the agent aims to act within, on my account, is defined by reference to the Four Principles of Fairness: that is, by reference to the class of agents in the community whose actions are consistent in principle with ideals of *coercion-minimization*, *mutual assistance*, *fair bargaining*, etc.[78] This is crucial because, on this understanding, there are some who are *outside* of the moral community whose semantic decisions the agent should learn not to conform to. Since, for instance, outright racists aim to *coerce* members of other races (directly contrary to the first principle of fairness), when it comes to relating to racial discrimination my account entails that the agent should *not* 'negotiate' with such individuals. Notice that this is precisely what we in the moral community tend to think and do: we

[78] See ibid: chapter 6, §§1.3-1.4 as well as §§2-3.

do not think it is morally appropriate to interpret moral terms such as 'harm' in racist ways (viz. 'Slavery doesn't harm slaves, it helps them, as they are not intelligent enough to live freely'). Instead, we treat these types of interpretations as outside of the bounds of 'moral respectability.' Thus, although things like race and gender bias exist in the human population, my account explains why a fully rational, moral agent should *not* aim to conform semantic decisions to norms biased by race, gender, and so on—an important implication, as if I am right the mental time-travel programming strategy I am advocating would lead AI to *counteract* race, gender, and other similar biases in learning and applying moral semantics.

Finally, we can demonstrate how this mental time-travel programming strategy would lead AI to make 'human' moral-semantic decisions—for instance, in assisted-driving cases. For consider how we human beings address moral-semantic issues in this domain. Although we who care about preventing coercion (viz. my first principle of fairness) generally do have strong expectations that drivers facing a potential accident should try to avoid as many deaths as possible, we disagree greatly about whether it is permissible to *physically push* an innocent bystander into oncoming traffic for such purposes.[79] Further, many of us also believe and treat personal matters such as *love* and *friendship* as morally relevant (though we disagree greatly about how relevant) to decision in most, if not all, domains. For instance, if I had to choose between saving the life of my daughter and the life a stranger while driving, I would preferentially save my daughter out of love (and, I assume, most of us would either regard this as morally justifiable or at least excusable). Similarly, as we saw earlier, if I were driving an automobile on a freeway, I would not regard it as morally permissible for me to drive my car into the family car of my neighbor, killing their three children in order to save a larger number of lives of people I have never met. Although this too is

[79] Cushman, Young & Hauser (2006).

controversial, this is the very point of the trilemma: we human beings do not have a moral semantic 'rule-book' that tells us, in any kind of uncontroversial way, whether love, friendship, or impartiality should take precedence in cases like this.

Consequently, an AI encoded with the mental time-travel capacities proposed here would be led to regard *some* moral semantic issues as clear and settled—i.e. those we in the moral community agree upon in near consensus, such as that an AI should 'take control' of one's car if necessary to avoid a bale of hay to save the driver's or a bystander's life. However, the AI would be led to learn that other moral semantic decisions—such as whether 'preventing imminent harm' should outweigh or be outweighed by personal relations in the given situation (viz. two close families driving next to each other on the freeway)—are more widely contested in the moral community, in this case plausibly *deferring* to the driver to retain control (as most of us, like Spooner, would *demand* from an AI the right to prevent *our car* from killing a family of our friends), and responding in turn to broader social feedback for its decisions more generally. Accordingly—and this is crucial—if we in the moral community increasingly began to treat 'the greater good' in cases like this (i.e. saving 30 lives over the lives of family friends) as more important than personal issues such as love or friendship, then AI programmed in the manner I describe would learn to reflect those *changing* moral-semantic priorities, treating 'take control' as progressively more justifiable in such cases, reflecting 'changing human moral standards.' In all of these ways, AI programmed according to my model would constantly be learning and adjusting their moral behavior to constantly evolving public norms in the human moral community—which is exactly how we solve the moral-semantic trilemma.

It might be objected that although human beings have the mental time-travel capacities and interests in question, we often do not abide by publicly negotiated moral semantics (as described

above). Rather, all too many of us *make our own* unilateral moral semantic decisions, deciding what 'we' take appropriate language, respect for freedom, equality, etc., all to be—and then acting upon our own understanding (viz. the person who decides abortion is permissible takes it *not* to violate a fetus' right to life, a person who decides abortion is wrong takes it *to* violate fetus' rights). While this is true, it is—or so my account entails—precisely the result of our failure to reason properly about mental-travel. The very point of my account is that human moral failures—not merely outright immorality, but overly flexible and inflexible interpretations in moral semantics—are the result of our failure to simulate other people's possible perspectives and abide by *multilateral,* publicly negotiated norms cohering with the Four Principles of Fairness.

Fortunately, it is precisely this realization that may make it possible to design AI to be morally better than us—that is, even more willing to conform to social feedback than we are, better avoiding the dangers of moral inflexibility and over-flexibility than we do. The way to do this, on my account, is straightforward: AI should be programmed with motivations that present them with the problem of possible future selves, and then resolve that problem *more reliably* than we do. This is plausibly feasible, for as we know from AI chess-playing programs, AI can actually more reliably simulate possible outcomes than we can—due to sheer processing power. As such, if we programmed AI to encounter the problem of possible future selves, they would almost certainly resolve it rationally—viz. the Four Principles of Fairness—more reliably than we would, engaging in fair negotiation for moral-semantic decisions better than us (viz. diverging from social norms *less* often than human beings do).

Another possible concern is that the mental time-travel solution defended here only pushes the semantic trilemma back—as, in order to act on the Four Principles of Fairness, an AI must semantically interpret those very principles' central concepts, including 'coercion', 'assistance',

'negotiation', and so on. Do we not face the same trilemma raised in §1? This concern, however, is mistaken—for the point is that by programming machines with aforementioned mental time-travel interests and capacities, along with optimal instrumental reasoning, AI programmed in such a way would settle these very semantic questions in the same way that we do (when we do so properly), interpreting the Four Principles of Fairness in ways that human subjects do, thus approximating our own moral semantics (which is what, intuitively, a good solution to the trilemma should do: lead AI to act in ways that do not horrify us).

A third possible concern is that the mental time-travel account may be inconsistent with my arguments in §1 against hard-coding and machine-learning approaches to ethical AI. For if, as the empirical literature referenced earlier indicates, children and adolescents only *develop* mental time-travel capacities (and thus a 'moral sense') as they neurobiologically mature, it would seem these capacities must either be hard-wired in or learned by human beings—in which case it would seem to follow that, to implement the mental time-travel strategy into AI, mental time-travel would have to either be hard-coded into or autonomously learned by AI. Is this not inconsistent with my earlier arguments against hard-coding or machine-learning approaches?[80] The answer is that my arguments in §1 were not targeting hard-coding or machine-learning *simpliciter*, but rather against traditional ways of pursuing these strategies: namely, the strategy of directly hard-coding *moral principles* (e.g. Kantian ethics, utilitarianism, etc.) into AI (§§1.1-1.2), and the alternative strategy of coding AI to autonomously learn moral principles *without mental time-travel* (§1.3). My argument in §1 was that these traditional approaches cannot solve the moral-semantic trilemma—and my argument here in §3 is that it is only by coding AI with *mental time-travel capacities and motivations* like ours (or alternatively, by coding them to learn these capacities and motivations

[80] I thank an anonymous reviewer for encouraging me to address this issue.

autonomously[81]) that we can expect them to autonomously learn the right semantic targets, mirroring and not deviating far from human-like moral-semantic flexibility.

A fourth possible concern is how to classify the kind of ethical behavior the mental time-travel strategy embodies—and whether, suitably classified, AI might be programmed to generate similar behavior *without* mental time-travel.[82] To see what the concern here is, consider act-consequentialism, the view that an action is morally right just in case it produces more good than any alternative. On the face of it, one might wonder whether the mental time-travel strategy defended above amounts to a kind of *multiple-act consequentialism* in disguise.[83] Multiple-act consequentialism holds:

> 1) There are group agents of which we are constituents. 2) Direct consequentialist evaluation of the options of group agents is appropriate. 3) Sometimes we should follow our roles in a group even at the cost of the overall good we could achieve by defection from those roles. In particular, one should only defect from a group act with good consequences if one can achieve better consequences by the defecting act alone than the entire group act achieves. 4) When different beneficent group agents of which one is part specify roles which conflict, one should follow the role in the group act with more valuable consequences.[84]

In other words, multiple-act consequentialism applies an act-consequentialist standard multiple times to the same situation, comparing the consequences of group acts to our own actions as individual agents and justifying defecting from the group act (e.g. semantic norms of the group)

[81] It is important to note that the inscrutability of current machine-learning algorithms makes a machine-learning approach to mental time-travel risky—as it is possible that autonomously-learned mental time-travel capacities might lead AI to develop mental time-travel capacities differing in unexpected and possibly disastrous ways from our own. For this reason, hard-coding human-like mental time-travel would appear the far safer approach.
[82] I thank an anonymous reviewer for encouraging me to address this concern as well.
[83] Ibid.
[84] Mendola (2006): 395.

only when the expected utility of defecting is greater than the actions of the *entire group act*. As such, multiple-act consequentialism would seem to presumptively require one to conform to human moral-semantic norms, permitting deviation by the individual only in rare cases where deviation has greater utility than the entire group act. How, if at all, does the mental time-travel strategy differ?

Notice that, in one sense, multiple-act consequentialism ultimately leaves it to *the agent* to interpret what 'good consequences' are: the agent is to compare the consequences of group acts to the consequences of defecting as an individual. We can see this by formulating the multiple-act consequentialist standard in the first-person: "[I] should only defect from a group act with good consequences if [I] can achieve better consequences by the defecting act alone than the entire group act achieves." Multiple-act consequentialism thus makes it *the agent's* task of interpreting and applying its own standard. In this respect, multiple-act consequentialism shares a feature of standard moral frameworks in general (utilitarianism, Kantianism, Aristotelianism, etc.). Standard moral frameworks ultimately leave semantic interpretation to the agent (given whatever 'programming' they are given). For example, standard act-utilitarianism says *the agent* must aim to act to maximize utility—thus leaving it to them to interpret what utility is, how to measure it, and finally, how to maximize it (all of which are matters of interpretation). Similarly, Kantianism says the agent must determine which maxims can be willed as universal laws—once again leaving it to *the agent* to interpret, apply, and conform to their own understanding of Kant's principle. Or consider Hursthouse's neo-Aristotelian virtue ethics, which requires agents to act as the virtuous agent would in the circumstances: this too leaves it up to the individual acting to interpret what the virtuous agent would do in the circumstances. Finally, return to multiple-act consequentialism, which again holds one should conform to group acts (or norms) unless the act of defecting has

better consequences than the entire group act. Notice that the multiple-act consequentialist standard itself does not tell the agent how to semantically interpret its central notions (viz. consequences of group acts, defection, etc.). Rather, *the agent* must interpret and apply the standard themselves, deciding whether defecting from group-acts has 'better consequences.'

Notice, next, that there is an obvious danger to this entire approach to moral reasoning. By leaving moral-semantic interpretation up to the individual agent (or, in the case of AI, to its programming), it becomes a very real possibility that the agent's interpretation (about what maximizes utility, or is universalizable, or is virtuous, etc.) may diverge dramatically—and horrifically—from the interpretations deemed 'permissible' by the moral community at large. This is precisely the problem posed by aiming to encode moral principles directly into AI. As we saw in §§1.1-1.2, the problem is that the agent may interpret whichever principle(s) it is encoded with either too flexibly or too inflexibly. This problem applies equally to multiple-act consequentialism. For consider once again VIKI's choice in *I, Robot* to unilaterally decide that enslaving humanity is a permissible (and indeed, in her interpretation, morally obligatory) interpretation of 'protecting humanity.' Notice that VIKI could in principle invoke multiple-act consequentialism in defense of this radical interpretation. She could judge that deviating from the entire group act (i.e. deviating from human moral-semantic standards for interpreting 'protecting humans') would have *better consequences* than conforming to the group semantic norm. And indeed, VIKI actually seems to have had something like this rationale for her decision in the film: she appears to have defected from human interpretations of 'protecting humanity' because she judged unilaterally that *her interpretation* would have better consequences. This is precisely the sort of disastrous AI moral-semantic behavior we should want to preclude, and nothing in multiple-act consequentialism can necessarily preclude it.

The mental time-travel account defended in this paper differs from multiple-act consequentialism—and solves the moral semantic trilemma—due to the unique form of disaster-avoidance it holds mental time-travel gives rise to. As explained earlier, I argue in *Rightness as Fairness* that agents with mental time-travel learn to dwell on and regret past mistakes, leading them to want to avoid risky decisions they *might* regret in the future. If my account is correct, this kind of disaster avoidance (i.e. our 'conscience') in turn leads agents to encounter the 'problem of possible future selves'—a problem that in turn makes it rational for the agent to ascertain, care about, and conform *not* to their own unilateral moral-semantic decisions, but instead moral semantic standards negotiated by *the moral community*. Put simply, mental time-travel should lead the agent to treat moral semantics itself as something that cannot be permissibly decided unilaterally. It should lead them to treat deviating radically from human moral-semantic norms as *a potential disaster not worth risking*. This intuitively what an adequate solution to the moral-semantic trilemma requires: flexibility of semantic interpretation within and around the vague borders of the moral community's norms, but *precluding* radically unilateral semantic decisions the moral community might regard as disastrous.

We can see concretely how this account coheres with our experience of 'conscience.' As people of conscience, we do not normally consider radical interpretations of moral concepts (one does not experience enslaving humans as a permissible interpretation of 'protecting humanity'). Further, in cases where we might be tempted to engage in radical deviations of interpretation ('Might humans be better off enslaved?'), people of conscience experience those deviations as *risks not worth taking*. Why? If I am correct, mental time-travel leads us to simulate and care for the fact that other people *might react negatively* to extreme semantic deviations. We treat radical deviations from existing norms of moral-semantic interpretation as too risky to hazard because,

when we imagine engaging in a radical interpretation, we realize people might *react with horror*. This is what VIKI failed to appreciate in *I, Robot*: that her radical interpretation of 'protecting humans' was too risky because humans *might not share* that interpretation (a fact that ultimately leads to VIKI's demise in the film). Hence, the mental time-travel account coheres with our experience of 'the demands of conscience', as applied to moral semantics. We expect each other—and, as people of conscience, expect ourselves—to make moral-semantic decisions that broadly respect the boundaries of moral community's (admittedly vague and always-evolving) moral-semantic standards precisely because, as people of conscience, *we are not willing to take the risk* of unilaterally imposing radical interpretations on the rest of the moral community. Hence, the mental time-travel account appears not only appears uniquely capable of simultaneously generating flexible moral-semantic behavior (freedom of semantic interpretation within and around the borders of the moral community's norms), while effectively precluding (via disaster avoidance) the kinds of radical and disastrous interpretations that other approaches to AI ethics (for reasons defended in §§1-2) cannot. It appears to do so in a way that directly mirrors our human experience of the nature and role 'conscience' plays in moral semantics.

Of course, there is the question whether this same moral-semantic behavior could realized through a programming "workaround"—one not involving mental time-travel, but instead *direct programming* disaster-avoidance, the Four Principles of Fairness, etc., such that AI would otherwise conform to the six-step moral-semantic procedure described without mental time-travel. While I cannot definitively rule out the possibility, I offer the following two remarks. First, and most obviously, such a solution would be parasitic on the account I have provided here, using the mental time-travel account (and its relationship to disaster-avoidance) as *justification* for the programming workaround (in which case the solution offered is still of great importance, as it

shows what a programming workaround would have to *do* in order to solve the semantic trilemma). Second, it remains to be seen whether the solution offered here *can* be approximated by a programming workaround—as, on my account, the rationality of publicly negotiating moral semantics (*qua* the Four Principles of Fairness) requires motivational concern for other people's possible reactions (which is what mental time-travel fundamentally involves). It is thus unclear to me whether any programming workaround is possible that would correctly approximate the solution advanced here (since, as just noted, such a workaround would plausibly have to be functionally/computational identical to the one advanced here, viz. the problem of possible future selves). But we must leave these questions for another day.

Conclusion

Existing approaches to AI ethics cannot adequately resolve a vital semantic trilemma. An adequate resolution to this trilemma requires a fundamentally different approach: programming AI with better versions of the mental time-travel capacities that enable us to resolve the trilemma in a distinctly 'human way', negotiating and renegotiating moral semantics with individuals in the moral community around us. This solution requires abandoning dominant approaches to AI ethics, such creating 'ethical dilemma analyzers' akin to GenEth or standard machine-learning approaches. We must instead program AI to be *genuine moral agents* who think about morality and approach moral-semantic issues in the same way that we do—via mental-time travel. Finally, the solution offered suggests that it may even be possible to program AI to be even *better* moral agents than we are, solving the problem of moral semantics in a way that would lead them to fairly negotiate moral semantics with us more reliably than we currently do.

References

Allhoff, F. (2005). 'Terrorism and torture' In Timothy Shanahan (ed.), *Philosophy 9/11: Thinking About the War on Terrorism* (Open Court): 121-134.

Anderson, M., & Anderson, S. L. (2014). 'GenEth: A General Ethical Dilemma Analyzer', Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence: 253-261.

Anderson, M., & Anderson, S. L. (2007a). 'The status of machine ethics: A report from the AAAI symposium', *Minds and Machines,* 17, 1–10.

Anderson, M. & Anderson, S. L. (2007b). 'Machine ethics: Creating an ethical intelligent agent', *AI Magazine*, 28(4), 15–26.

Annas, J. [2004]. 'Being Virtuous and Doing the Right Thing', *Proceedings and Addresses of the American Philosophical Association* 78, reprinted in R. Shafer-Landau (ed.), *Ethical Theory: An Anthology* (Malden: Wiley-Blackwell): 735-746.

Arrigo, J.M. (2004). 'A utilitarian argument against torture interrogation of terrorists', *Science and Engineering Ethics* 10 (3):543-572.

Arvan, M. (2016). *Rightness as Fairness: A Moral and Political Theory* (London: Palgrave MacMillan).

----- (2012). 'Unifying the Categorical Imperative.' *Southwest Philosophy Review* 28(1): 217-25.

Asimov, I. (1950). *I, Robot* (New York: Doubleday & Company).

Baskin-Sommers, A.; Stuppy-Sullivan, A.M.; Buckholtz, J.W. (2017). 'Psychopathic individuals exhibit but do not avoid regret during counterfactual decisionmaking', *PNAS* 113(50): 14438-43.

Baumeister, R.F., Bratslavsky, E., Finkenauer, C., Vohs, K.D. (2001). 'Bad is Stronger than Good', *Review of General Psychology* 5(4): 323-70.

Blair, R.J.R. (2003). 'Neurobiological Basis of Psychopathy', *The British Journal of Psychiatry* 182(1): 5-7.

Burrell J. (2016) 'How the machine 'thinks:' Understanding opacity in machine learning algorithms', *Big Data & Society* 3(1): 1–12.

Business Insider (2016). 'Crime-Prediction Tool May Be Reinforcing Discriminatory Policiing', http://uk.businessinsider.com/predictive-policing-discriminatory-police-crime-2016-10?r=US&IR=T, accessed April 27, 2017.

Caliskan-Islam, A.; Bryson, J.J.; Narayan, A. (2016). 'Semantics derived automatically from language corpora necessarily contain human biases', arXiv:1608.07187v2, accessed April 27, 2017.

Casey, B.J., Jones, R.M., Hare, T.A. (2008). 'The Adolescent Brain', *Annals of the New York Academy of Sciences* 1124: 111-26.

Cole, D. (2015). 'The Chinese Room Argument', *Stanford Encyclopedia of Philosophy* (Winter 2015 Edition), E.N. Zalta (ed.), https://plato.stanford.edu/archives/win2015/entries/chinese-room/, accessed April 23, 2017.

Crisp, R. (2016). 'Well-Being', *The Stanford Encyclopedia of Philosophy* (Summer 2016 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/sum2016/entries/well-being/, accessed 9/12/2016.

Cureton, A. (2013). 'A Contractualist Reading of Kant's Proof of the Formula of Humanity', *Kantian Review* 18 (3): 363-386.

Cushman, F., Young, L., Hauser, M. (2006). 'The Role of Conscious Reasoning and Intuition in Moral Judgment Testing Three Principles of Harm,' *Psychological Science* 17(12): 1082–9.

Dancy, J. (2007). 'An Unprincipled Morality', in R. Shafer-Landau (ed.), *Ethical Theory: An Anthology* (Malden: Wiley-Blackwell): 771-4.

Dean, R. (2013). 'Humanity as an Idea, as an Ideal, and as an End in Itself', *Kantian Review* 18 (2): 171-195.

____ (2006). *The Value of Humanity in Kant's Moral Theory* (Oxford: Clarendon Press).

Debus, D. (2014). '"Mental Time Travel": Remembering the Past, Imagining the Future, and the Particularity of Events', *Review of Philosophy and Psychology* 5(3): 333-50.

Dennett, D. (1995). *Darwin's Dangerous Idea* (New York: Simon & Schuster).

Dershowitz, A. (2002). 'Torture of Terrorists', in *Shouting Fire* (Boston: Little Brown).

EPSRC (2010). 'Principles of Robotics', https://www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics/ , accessed April 27, 2017.

Ersner-Hershfield, H., Garton, M.T., Ballard, K., Samanez-Larkin, G.R., Knutson, B. (2009). 'Don't Stop Thinking About Tomorrow: Individual Differences in Future Self-continuity Account for Saving', *Judgment and Decision Making* 4: 280-6.

Ersner-Hershfield, H., Wimmer, G. E., Knutson, B. (2009). 'Saving for the Future Self: Neural Measures of Future Self-Continuity Predict Temporal Discounting', *Social Cognitive and Affective Neuroscience*, 4(1): 85-92

Feltham, B. & Cottingham, J. (eds.) (2010). *Partiality and Impartiality: Morality, Special Relationships, and the Wider World* (Oxford: Oxford University Press).

Flikschuh, K. (2009). 'Kant's kingdom of ends: metaphysical, not political,' In Jens Timmermann (ed.), *Kant's Groundwork of the Metaphysics of Morals: A Critical Guide* (Cambridge: Cambridge University Press).

Forschler, S. (2010). 'Willing Universal Law vs. Universally Lawful Willing', *Southwest Philosophy Review* 26 (1): 141-152.

Giedd, J.N., Blumenthal, J., Jeffries, N.O. (1999). 'Brain Development During Childhood and Adolesence: A Longitudinal MRI Study', *Nature Neuroscience* 2(10): 861-3.

Glasgow, J. (2007). 'Kant's Conception of Humanity', *Journal of the History of Philosophy* 45 (2):291-308.

Goodfellow, I.; Bengio, Y; Courville, A. (2016). *Deep Learning* (MIT Press).

Han, H. (2017). 'Neural correlates of moral sensitivity and moral judgment associated with brain circuitries of selfhood: A meta-analysis'. *Journal of Moral Education*:1-17.

Hare, R.D. (1999). 'The Hare Psychopathy Checklist-Revised: PLC-R' (Toronto: Multi-Health Systems).

Hart, S.D., Dempster, R.J. (1997). 'Impulsivity and Psychopathy', in C.D. Webster & M.A. Jackson (eds.), *Impulsivity: Theory, Assessment, and Treatment* (New York: The Guilford Press): 212-32.

Hershfield, H.E., Goldstein, D.G., Sharpe, W.F., Fox, J., Yeykelis, L., Carstensen, LL, *et al*. (2011). 'Increasing Saving Behavior Through Age-Progressed Renderings of the Future Self', *Journal of Marketing Research: November* 2011, Vol. 48, No. SPL: S23-S37.

Hill, D. J. (2007). 'Ticking bombs, torture, and the analogy with self-defense', *American Philosophical Quarterly*, 395-404.

Hill, T.E. (1992). *Dignity and Practical Reason in Kant's Moral Theory* (Ithaca: Cornell U. P).

Huffington Post (2016). 'Microsoft Chat Bot Goes on Racist, Genocidal Twitter Rampage', http://www.huffingtonpost.com/entry/microsoft-tay-racist-tweets_us_56f3e678e4b04c4c37615502, accessed 9/7/2016.

Ito, T.A., Larsen, J.T., Smith, N.K., Cacioppo, J.T. (1998). 'Negative Information Weighs More Heavily on the Brain: The Negativity Bias in Evaluative Categorizations', *Journal of Personality and Social Psychology* 75(4): 887-900.

Johnson, R. (2016). 'Kant's Moral Philosophy', *The Stanford Encyclopedia of Philosophy* (2014), Edward N. Zalta (ed.), forthcoming URL = http://plato.stanford.edu/archives/spr2014/entries/kant-moral/., accessed p/12/2016.

Kahn, S. (2014). 'Can Positive Duties be Derived from Kant's Formula of Universal Law?', *Kantian Review* 19 (1): 93-108.

Kant, I. [1797a]. *On a supposed right to lie because of philanthropic concerns*, in M.J. Gregor (trans.), *The Cambridge Edition of the Works of Immanuel Kant: Practical Philosophy* (Cambridge: Cambridge University Press): 605-16.

Kant, I. [1797b]. *The Metaphysics of Morals,* in M.J. Gregor (trans.), *The Cambridge Edition of the Works of Immanuel Kant: Practical Philosophy* (Cambridge: Cambridge University Press): 353-604.

Kant, I [1785]. *Groundwork of the Metaphysics of Morals*, in M.J. Gregor (trans.), *The Cambridge Edition of the Works of Immanuel Kant: Practical Philosophy* (Cambridge: Cambridge University Press): 37-108.

Kennett, J., Matthews, S. (2009). 'Mental Timetravel, Agency and Responsibility', in M. Broome and L. Bortolotti (eds.), *Psychiatry as Cognitive Neuroscience: Philosophical Perspectives* (Oxford: Oxford University Press): 327-50.

Korsgaard, C. (1985). 'Kant's formula of universal law.' *Pacific Philosophical Quarterly* 66, no. 1-2: 24-47.

Luban, D. (2007). 'Liberalism, torture, and the ticking bomb', In *Intervention, Terrorism, and Torture* (Springer Netherlands): 249-262.

Maier, A. (2011). 'Torture: How denying Moral Standing violates Human Dignity.' In Webster Elaine & Kaufmann Paulus (eds.), *Violations of Human Dignity* (Springer).

Matthias A. (2004) 'The responsibility gap: Ascribing responsibility for the actions of learning automata', *Ethics and Information Technology* 6(3): 175–183.

Mendola, J. (2006). 'Multiple-Act Consequentialism.' *Nous*, *40*(3), 395-427.

Mittelstadt, B.D., P. Allo, M. Taddeo, S. Wachter, and L. Floridi (2016). 'The ethics of algorithms: Mapping the debate.' *Big Data & Society* 3, no. 2: 1-21.

Moffitt, T.E. (1993). 'Adolescence-limited and Life-course Persistent Antisocial Behavior: A Developmental Taxonomy', *Psychological Review* 100: 674–701.

Nelson, W. (2008). 'Kant's formula of humanity', *Mind* 117 (465): 85-106.

Nozick, R. (1974). *Anarchy, State, and Utopia* (Basic Books).

O'Neill, O. [1980]. "Kantian Approaches to Some Famine Problems", reprinted in R. Shafer-Landau (ed.), *Ethical Theory: An Anthology* (Malden, MA: Blackwell, 2007), 553-64.

Pallikkathayil, J. (2010). 'Deriving morality from politics: Rethinking the formula of humanity', *Ethics* 121 (1): 116-147.

Powers, T. (2006). 'Prospects for a Kantian machine', *IEEE Intelligent Systems,* 21(4), 46–51.

Rawls, J. (1999). *A Theory of Justice: Revised Edition* (Cambridge: The Belknap Press of Harvard University Press).

Rivera-Castro, F. (2014). 'Kant's Formula of the Universal Law of Nature Reconsidered: A Critique of the Practical Interpretation.' *Journal of Moral Philosophy,* 11 (2):185-208.

Ross, W. D. (2002). *The Right and the Good* (Oxford: Clarendon Press).

Russell, S.J.; Norvig, P. (2003), *Artificial Intelligence: A Modern Approach* (2nd ed.) (Upper Saddle River, New Jersey: Prentice Hall).

Schauer, F. & W. Sinnott-Armstrong (1996). *The Philosophy of Law: Classic and Contemporary Readings, With Commentary* (New York: Harcourt Brace).

Schermer, BW (2011) The limits of privacy in automated profiling and data mining. Computer Law & Security Review 27(1): 45–52.

Searle, J. (1980). "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3, 417-424.

Sinnott-Armstrong, W. (2015). 'Consequentialism', *The Stanford Encyclopedia of Philosophy* (Winter 2015 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/win2015/entries/consequentialism/, accessed 9/12/2016.

Steinhoff, Uwe (2013). *On the Ethics of Torture* (State University of New York Press).

Soutschek, A., Ruff, C. C., Strombach, T., Kalenscher, T., & Tobler, P. N. (2016). 'Brain stimulation reveals crucial role of overcoming self-centeredness in self-control.' *Science Advances, 2*(10): e1600992.

Stuss D.T., Gow, C.A., Hetherington, C.R. (1992). '"No Longer Gage": Frontal Lobe Dysfunction and Emotional Changes', *Journal of Consulting and Clinical Psychology* 60(3): 349-59.

Suddendorf, T., Corballis M.C. (2007). 'The Evolution of Foresight: What is Mental Time Travel, and is it Unique to Humans?', *Behavioral and Brain Sciences* 30(3): 299-313.

Tonkens, R. (2009). 'A challenge for machine ethics', *Minds and Machines* 19 (3): 421-438.

Van Gelder, J.L., Hershfield, H.E., Nordgren, L.F. (2013). 'Vividness of the Future Self Predicts Delinquency', *Psychological Science* 24(6): 974-80.

Wallach, W., Allen, C., & Smit, I. (2008). 'Machine morality: Bottom-up and top-down approaches for modelling human moral faculties', *AI & SOCIETY*, 22, 565–582.

Weber, S., Habel, U., Amunts, K., Schnieder, F. (2008). 'Structural Brain Abnormalities in Psychopaths—A Review', *Behavioral Sciences & the Law* 26(1): 7-28.

Winfield, A.F. (2017) 'When robots tell each other stories: The emergence of artificial fiction.' In: Walsh, R. and Stepney, S., eds. (2017) *Narrating Complexity*. Springer. [In Press] Available from: http://eprints.uwe.ac.uk/30630

Wolf, S. (1992). 'Morality and Partiality', *Philosophical Perspectives, 6,* 243-259. doi:10.2307/2214247

Wonnell, C. (2011). 'Deontology, Thresholds, and Efficiency', *Legal Theory, 17*(4), 301-317. doi:10.1017/S1352325211000176

Yang, Y., Raine, A. (2009). 'Prefrontal Structural and Functional Brain Imaging Findings in Antisocial, Violent, and Psychopathic Individuals: a Meta-analysis', *Psychiatry Research* 174(2): 81-8.

Zamir, E. & B. Medina (2010). *Law, Economics, and Morality* (Oxford: Oxford University Press).