The Importance of Metaethics for Artificial Moral Intelligence

by

David Gantt

Honors Thesis

Presented to the Department of Philosophy

Tufts University

Medford, Massachusetts

Advisor: Professor Patrick Forber

Committee Members: Professor Sigrún Svavarsdóttir, Professor Brian Epstein

May 13th, 2022

Acknowledgments

I am grateful to the following people for their feedback, inspiration, and support: Professor Patrick Forber, for serving as both my major and honors thesis advisor and for being an invaluable and generous resource and mentor; Professor Sigrun Svavarsdottir and Professor Brian Epstein, for serving as committee members, recommending readings, and answering my questions during the writing process; my brothers, Will and Ben, for providing meticulous feedback and, in part, inspiring my interest in these topics. It is only by the guidance of these people that this project has come to fruition.

Table of Contents

Introduction

On October 14th, 2021, the Allen Institute for AI released Delphi, a machine learning model that makes normative judgments about social and moral scenarios. When one inputs a description of or question about such a scenario—e.g., "robbing a bank if you are poor" or "is it OK to get an abortion?"—, Delphi outputs a judgment. The model says that ignoring a phone call from your friend is "rude," cleaning a toilet bowl with your shirt is "disgusting," and parking in a handicap spot when you are not disabled is "wrong." Impressively, it distinguishes between similar scenarios: for example, it says that it is wrong not to pay attention in class but that it is understandable if you have ADHD. Researchers at the Allen Institute say that Delphi "demonstrates strong promise of language-based commonsense moral reasoning" and makes these judgments with 92.1% accuracy.[1] This machine learning model represents the latest and most successful attempt at inculcating in artificial intelligence (AI) a distinctly human capacity: moral judgment.

Other research institutions have attempted similar projects. In 2014, MIT released the Moral Machine, an online platform on which people respond to moral dilemmas about autonomous driving. For example, users may be prompted to decide whether an autonomous vehicle should crash into a barrier, killing its passenger, or instead run over an elderly woman and her dog in the crosswalk. Users can see how their responses align with those of other users or create scenarios themselves. This experiment was quite successful: the platform gathered "40 million decisions in ten languages from millions of people in 233 countries and territories" and discovered different trends in responses from Western, Eastern, and South American countries.[2] The MIT researchers' aimed to democratize moral decision-making about the distribution of

[1] Liwei et al. "Delphi: Towards Machine Ethics and Norms." pp. 1 (2021).
[2] Awad et al. "The Moral Machine experiment." *Nature* 563, pp. 59 (2018).

well-being and harm by crowdsourcing moral judgments. In their words, "decisions about the ethical principles that will guide autonomous vehicles cannot be left solely to either the engineers or the ethicists."[3] Delphi and the Moral Machine are recent examples of an ambitious project in AI research: to create models that capture the nuances of human morality.

AI plays an increasingly pervasive and broad role in modern society. Critical tasks once performed by humans now fall under the province of AI. People trust autonomous cars to drive them, automated financial advisors to make their investment decisions, and robots to perform life-saving surgeries. Many experts foretell that researchers will develop artificial general superintelligence—"an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills"—by about 2050.[4] As AI systems perform more consequential tasks, there is rising concern about their inability to think in moral terms. AI will not only need to "promote well-being and minimize harm, but also to distribute the well-being they create, and the harm they cannot eliminate."[5] The creation of a moral AI—one which could learn moral norms and deliberate about moral decisions—would increase the chance that AI would be beneficent. Not only would an AI with moral understanding be less likely to accidentally harm humans, but it would be resistant to pernicious use by terrorists, rogue state actors, and criminals.

Any moral intelligence, be it human or artificial, must do four things: first, it must learn moral and social norms; second, it must perceive real-world situations, be it visually or textually; third, it must anticipate the outcome of alternative actions in these situations; and fourth, it must make ethical judgments given tradeoffs between values in different contexts.[6] Two of these

[3] Awad et al, "The Moral Machine experiment," pp. 59.
[4] Bostrom, Nick. "How Long Before Superintelligence?" (1997).
[5] Awad et al, "The Moral Machine experiment," pp. 59.
[6] Liwei et al, "Delphi: Towards Machine Ethics and Norms," pp. 1.

tasks, perception and planning, are technological problems, though they are informed by research in neuroscience and other fields. The other two tasks, implementing moral understanding and deliberation, are philosophically harder. They require answers to metaethical questions about the substance and acquisition of moral knowledge.

Researchers developing moral AI must therefore ask questions at the intersection of metaethics and machine learning: How does one, person or machine, acquire moral knowledge? If moral truths exist, are they relative to different moral communities, or are they universal? How can AI be updated to reflect a society's evolving moral convictions? Would AI be motivated by the variety of moral emotions people experience, like disgust, shame, and sympathy? Philosophers' disagreements on these questions are deep-seated. It will become clear that one's metaethical assumptions determine one's approach to designing a moral AI.

The goal of this paper is to lay out the tenets of three metaethical views—expressivism, non-natural realism, and natural realism—and discuss their implications for the design of moral machine learning. In doing so, I will also evaluate the degree to which artificial moral agency may be like human moral agency. I lay out metaethical problems for this project and in doing so aim to affirm the relevance of metaethics in the modern world. Moral machine learning is a case of a centuries-old, esoteric field raising issues for one of the most exciting developments in modern science. But, to this point, AI researchers have failed to be explicit about the metaethical assumptions they make.

This essay comprises four parts: First, I will introduce both metaethics and machine learning. Second, I will argue that artificial agents may mimic, but lack important features of, human moral agency. Third, I will survey common approaches to making these agents, distinguish between top-down and bottom-up approaches, and argue that the latter is the only

sensible approach. Fourth, I will outline the implications of three metaethical views for moral machine learning.

Background: Metaethics

Metaethics dates back millennia and machine learning only decades.[7] Philosophers pay little mind to machine learning, and technologists pay even less mind to metaethics, so it will be helpful to provide background knowledge on both topics. I do not survey common metaethical views in this section, which I will do later in the paper. Here, I provide a broad overview of the field.

Philosophers make a useful distinction between three subfields in ethics: normative ethics, applied ethics, and metaethics. Normative ethics first comes to most people's minds when they speak of ethics. It is concerned with determining which actions are right and wrong and by which criteria. Applied ethics, sometimes called practical ethics, is concerned with identifying and analyzing the ethical dimensions of real-world issues in medicine, law, technology, and other enterprises. Metaethics tries to make sense of the foundations of normative and applied ethics. It is concerned with the "metaphysical, epistemological, semantic, and psychological presuppositions and commitments of moral thought, talk, and practice."[8] Metaethicists tackle questions such as the following: Do moral judgments report truth, express sentiments, or reflect our interpersonal and societal conventions? How do we learn what is moral? Are moral standards culturally relative or universal? How does morality motivate us? And why do people disagree about moral issues so much? The landscape of metaethical views is diverse, and according to

[7] Plato's *Euthyphro* is maybe the first metaethical investigation. It asked what makes actions *pious*: "Is the pious loved by the gods because it is pious, or is it pious because it is loved by the gods?" A 1956 conference on AI at Dartmouth College is the event which founded the field of artificial intelligence.
[8] Sayre-McCord, Geoff, "Metaethics", *The Stanford Encyclopedia of Philosophy* (Summer 2014 Edition)

Australian philosopher Michael Smith, "there are no dominant views."[9] Some philosophers like Peter Railton believe that there is a moral reality independent of human beings, whereas others, like John Mackie, think that facts about morality do not exist. As is true in many philosophical disciplines, disagreements in metaethics have persisted for hundreds of years, and the number of respected views has proliferated.

The reason we care to debate and study normative ethics is clear: if nothing else, moral norms have social value. One's adherence to them is generally to everyone's social benefit, so moral transgressors risk paying great interpersonal, social, and economic costs. To avoid paying these costs, one must understand what it is right to do. Ethics is of interest to us by definition; typically, our moral judgments motivate us to act in accordance with them. By contrast, an interest in metaethical questions demands explanation. If people understand what behavior is ethical, why should we care about its foundations? After all, physicists carry on without any concern for metaphysics. Some moral philosophers have thought that resolution about metaethical questions is a precondition for engaging with normative ethical questions. If they are correct, metaethics would be disanalogous to metaphysics in this regard. However, most metaethicists reject this line of thinking. Indeed, the mainstream view is that metaethicists are engaged in an inquiry about moral thought, talk, and practice that is like the philosopher of science's inquiry about scientific thought, talk, and practice. Most metaethicists think that ethics should be pursued prior to metaethical resolution, much like most philosophers of science would encourage scientific inquiry. However, metaethics may nonetheless inform our understanding of how to engage in ethical debate, much as philosophy of science informs our understanding of how to engage in the scientific process. The value of metaethical inquiry lies at least in part in

[9] Smith, Michael. *The Moral Problem*. (Wiley-Blackwell, 1994), pp. 4.

this information. Moreover, I will argue, questions about moral learning and motivation are central to the task of creating moral AI.

Background: Machine Learning

People interact with machine learning every day without being aware of it. Machine learning determines the advertisements people look at, the roads they take to avoid traffic jams, which emails they see, and which go to their spam folder. Machine learning models invest their 401(k) funds, guide unmanned military drones, and can help identify lung cancers. Machine learning models are used in every sector, both public and private, and they pervade every corner of society where innovation is occurring.

AI is the field concerned with developing computer systems which perform tasks which require humanlike intelligence, like visual perception, linguistic skills, and moral decision-making. Machine learning is the subfield of AI which aims to accomplish these tasks by endowing computer systems with the ability to learn from experience—experience being, for a model, new data. After a period of training, during which a system adjusts its parameters—numerical variables internal to the model which guide its decision-making—it may perform a task without human guidance. Unlike many manufacturing machines, for example, machine learning models are not explicitly programmed to accomplish a specific task, making machine learning techniques well-suited for complex tasks for which there is no easy algorithmic solution. Thus far, most machine learning models are trained for a single purpose, like playing chess, driving cars, or classifying images. In other words, their intelligence is minimally transferable: a world-class AI chess player may lose to a child in checkers. The task of making a generally

intelligent model, one which outperforms humans in a variety of tasks, is top of mind for researchers today.

There is a distinction between machine learning algorithms and machine learning models*:* algorithms are the procedures that computers run to create a model, which is the agent doing the decision, prediction, or recommendation. There are three types of machine learning algorithms: supervised, unsupervised, and reinforcement learning. First, supervised learning is the technique by which a model learns a function to map inputs to outputs. This sort of learning is supervised in the sense that the model is trained on a human-labeled data set. Such a model might learn to classify images by the kind of food they contain. Second, unsupervised learning occurs when the system attempts to uncover patterns from unlabeled data. Such a model might learn what groups of consumers are interested in which products (in fact, this kind of algorithm is commonly used for personalized advertising). Last, reinforcement learning refers to teaching agents to complete complex objectives by rewarding certain behaviors and potentially punishing others. For example, autonomous vehicles may be rewarded for driving in their lanes, taking smooth turns, or avoiding busy routes.

What machine learning models thus far lack in generality, they make up for in speed. Because "the brain can perform at most about a thousand basic operations per second, or 10 million times slower than the computer," computer systems have surpassed humans on a range of tasks.[10] After a mere four hours of training, the model AlphaZero defeated world champions in the board games chess, shogi, and Go. Another model, AlphaFold, predicted the ways in which amino acids fold to compose proteins, a once intractable problem in biology. These models use

---

[10] Luo, Liqun. "Why Is the Human Brain so Efficient?" *Nautilus | Science Connected*, 9 Feb. 2022.

neural networks, a sort of artificial imitation of the human brain, to accomplish these tasks. Researchers speculate that, as the size of these models scale up, the pace of their learning will rapidly increase.

Thinkers like Nick Bostrom, Paul Christiano, and I.J. Good maintain the view that a transformative AI is not far afield. Although they profess some ignorance about the timeline of AI development, they estimate that researchers will develop "an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills" by about 2050.[11] However, there is disagreement about this prediction. In a recent survey about the developmental timeline of a transformative AI, 45% of experts predicted that it would occur before 2060, 34% predicted that it would occur after 2060, and 21% predicted it would never occur.[12] Patrick Winston, former director of the Computer Science and Artificial Intelligence Laboratory at MIT, remarked that predicting this date is difficult: "As far as a technological singularity, that requires one or more breakthroughs, and those are hard to think of in terms of timelines."[13] In any case, one should take these predictions with a grain of salt.

Although experts sometimes sound alarmist about the threat AI poses, there are indeed credible threats. The depictions in popular culture of AI gone awry (e.g., *2001: A Space Odyssey* or *Ex Machina*) seem to be the stuff of far-off science fiction, but AI systems are already malfunctioning in harmful ways. Whereas human intelligence is relatively efficient, robust, and generalizable, today's narrow, single-purpose AI is brittle. On May 6th, 2010, the S&P 500 crashed almost 10% in 36 minutes of trading in what is known as the "flash crash." Although the

[11] Bostrom, Nick. "How Long Before Superintelligence?" (1997).
[12] Dilmegani, Cem. "When Will Singularity Happen? 995 Experts' Opinions on AGI" (2017).
[13] Galeo, Dom. "Separating Science Fact from Science Hype: How Far off Is the Singularity?" *Futurism*, Futurism, 30 Jan. 2018.

precise cause of the crash is still debated, many believe that machine learning-based trading algorithms played an important role in precipitating the crash by irregularly selling large quantities of exchange-traded funds. In another case, a Tesla's autopilot system mistook the moon for a yellow traffic light and unexpectedly slowed down, imperiling its passengers. These cases illustrate that even narrow AI may have catastrophic effects. A general AI, although it does not exist yet, may pose an existential threat. A general AI in the hands of terrorists, a rogue state, or other malicious actor may be weaponized for "exploiting software vulnerabilities, for example…[or] generating political discord or misinformation with synthetic media; or initiating physical attacks using drones or automated weapons."[14] A general AI may also produce undesirable, unintended consequences in pursuit of a seemingly beneficent goal. Consider the following case from Stuart Russell:

> [Y]ou might ask the machine to counter the rapid acidification of the oceans that results from higher carbon dioxide levels. The machine develops a new catalyst that facilitates an incredibly rapid chemical reaction between ocean and atmosphere and restores the oceans' pH levels. Unfortunately, a quarter of the oxygen in the atmosphere is used up in the process leaving us to asphyxiate slowly and painfully. Oops.[15]

Many factors exacerbate the risk AI poses. For one, most AI innovation occurs in private labs and companies, making governmental oversight and regulation difficult. Note that AI is distinct in this regard from the development of, for example, nuclear technology, which occurred under government oversight in the United States. Profit-motivated companies are racing to develop a more robust and general AI, and they are incentivized to cut corners with respect to safety. Although the existential threat that AI may pose is decades or centuries away, narrow AI poses a threat commensurate with its increasing prevalence in society.

---

[14] Vold, Karina & Harris, Daniel R. (forthcoming). "How does Artificial Intelligence Pose an Existential Risk?" pp. 22-23, *Oxford Handbook of Digital Ethics*.
[15] Russell, Stuart. *Human Compatible: Artificial Intelligence and the Problem of Control*. pp. 138 (2019). Print.

As humans continue to cede morally consequential tasks to AI, it is imperative that we ensure that "these models capture our norms and values, understand what we mean or intend, and, above all do what we want."[16] One approach to ensuring that AI is beneficent is to develop in these models a moral intelligence like that of a human being. My goal in this essay is to lay out the implications of the metaethical assumptions researchers must make.

Artificial Moral Agency

Here, I argue that AI may mimic, but lacks important features of, human moral agency. AI may possess many important features of human moral agency: the capacities to apprehend moral norms, pass moral judgments, and justify those judgments. Moreover, they may mimic humanlike intentionality and motivation. But mimicking moral agency is not the same as being a moral agent. AI lack important features which are central to human moral agency, like subjective moral emotions which motivate moral action, human biological and social drives, and recognition of moral norms as such. Here, I define moral intelligence as the capacity to know right from wrong, weigh competing values, justify one's moral judgments, and exhibit moral emotions like disgust, sympathy, and compassion. Moral agents are those who possess this intelligence and act in accordance with it. Most humans, except infant children and psychopaths, are moral agents. To a lesser extent, so are some mammals like chimpanzees and capuchins. The question arises: can AI be moral agents in the way humans are? If not, to what extent can it mimic human moral agency?

Ethical agency may be more or less sophisticated. Philosopher James Moor distinguishes between three types of ethical agents: implicit, explicit, and full ethical agents. (Note that only

[16] Christian, Brian. *The Alignment Problem.* pp. 20 (WW Norton & Company: 2020). Print.

full ethical agents qualify as moral agents on my definition above.) Moor says that a machine is an implicit ethical agent "when the machine's construction addresses safety or critical reliability concerns."[17] These agents necessarily possess the ability to make decisions without human input, but they do not refer to ethical rules. For example, the autopilot system in a plane or other autonomous vehicle is an implicit ethical agent. It is ethically consequential that the autopilot delivers passengers safely and promptly to their destination, yet the autopilot need not refer to a normative theory to do this. Automated credit scorers, mortgage lenders, and investors are also implicit ethical agents. These agents are never held to account for their actions; rather, their manufacturers or irresponsible users are.[18] Tesla, for example, faced a lawsuit in 2018 for an autonomous vehicle which killed a pedestrian. Boeing faced similar legal issues after a malfunctioning auto-throttle led to a 737 crash. Although autopilots possess agency in the minimal sense that they make independent and self-initiated decisions, they do not exhibit humanlike intentionality, self-awareness, and self-control, so people do not hold autopilots morally accountable. For this same reason, one does not praise an autopilot when it functions well. Often, we attribute the failure to the implicit ethical agent but hold its manufacturers or irresponsible users to account.

An explicit ethical agent is one in which "ethics exist explicitly in [the] machine."[19] These agents can deal with ethical rules. To create such a machine, programmers formalize and implement theories of, for example, deontic or utilitarian logic. Unlike implicit ethical agents, examples of explicit ethical agents (both biological and artificial) are hard to come by. An example may be a sort of utilitarian algorithm, into which one inputs various utility functions,

[17] Moor, James (2009). "Four Kinds of Ethical Robots". *Philosophy Now* pp. 72:12-14.
[18] Zimmermann, Annette & Lee-Stronach, Chad (2021). Proceed with Caution. *Canadian Journal of Philosophy*.
[19] Moor, "Four Kinds of Ethical Robots"

probabilities, durations, etc., and the algorithm returns a utility-maximizing course of action. Moor argues that artificial explicit ethical agents would be well-suited to respond to natural disasters, during which "humans often have difficulty tracking and processing information about who needs the most help and where they might find effective relief."[20] One can imagine an artificial explicit ethical agent triaging hospital patients based on the urgency of their care. Explicit ethical agents determine their course of action by explicit reference to normative theories, but they do not satisfy any more conditions for moral responsibility than implicit ethical agents. Like implicit agents, explicit agents are not intelligent, and the fact that they encode ethical principles does not change this. There is therefore no reason to praise, blame, or hold them more accountable than implicit ethical agents.

Full ethical agents "make explicit ethical judgments and generally [are] competent to reasonably justify them."[21] Moreover, full ethical agents must possess mental states thought to be present only in intelligent animals, like beliefs, desires, intentions, and self-awareness. If any agent may be held morally responsible, it is a full ethical agent. They satisfy necessary conditions for moral responsibility, like intentionality, self-awareness, and self-control. Full ethical agents therefore deserve praise and blame for their actions.

It is a subject of debate whether an artificial system could meet the cognitive criteria for full ethical agency. That is, it is unclear whether computers may exhibit mental states like belief, doubt, or understanding. The philosopher John Searle, for example, criticizes the idea that an AI can "literally be said to understand [a] story."[22] But Searle thinks this will change: "The question, 'Can you build an artificial machine that is conscious?' is just like the question 'Can

[20] Moor, "Four Kinds of Ethical Robots"
[21] Moor, "Four Kinds of Ethical Robots"
[22] Searle, John. "Minds, Brains, and Programs". *The Behavior and Brain Sciences*. (1980) 3, pp. 417-457

you build an artificial heart that pumps blood?' We know how to build artificial hearts because we know how the biological heart works…assuming we knew how the brain worked, I see no obstacle in principle to building an artificial conscious machine."[23] Others demur. The computer scientist Subhash Kak says, "brains integrate and compress multiple components of an experience, including sight and smell—which simply can't be handled in the way today's computers sense, process and store data."[24] The question of whether AI may experience subjective mental states is likely to remain open for some time, and it is unclear what evidence may be presented that would resolve it. Yet, an AI can certainly mimic human belief, doubt, understanding, intentionality, and self-awareness. A neural network may be said to "believe" or "doubt" that a certain image depicts a cat; a neural language model may be said to "understand" a story if it can summarize it; an AI may be "intentional" insofar as it acts predictably in pursuit of goals (I examine this more closely below); and a robot may be "self-aware" by understanding its position relative to its environment. AI may mimic human belief, doubt, understanding, intentionality, and self-awareness, whether or not it feels these things subjectively.

I said in the introduction that the ability to apprehend moral norms is a necessary feature of moral intelligence. However, one may object that moral AI do not apprehend norms as such—that is, although moral AI render humanlike moral judgments, they do not recognize or refer to norms in making those judgments. This is a reasonable concern, and it leads one to wonder: is it possible to possess moral intelligence without understanding moral norms as such? A multi-layered, black-box neural network refines its weights during training to minimize a loss function. There is no sense in which the model possesses a notion of a norm or understands moral scenarios by applying moral norms. Rather, AI operates the way a moral particularist depicts

[23] Turello, Dan. "Brain, Mind, and Consciousness: A Conversation with Philosopher John Searle". (2015)
[24] Kak, Subhash. "Why a Computer Will Never Be Truly Conscious." *Governing*, Governing, 21 Apr. 2021..

moral thought and agency. AI does not operate by applying general moral norms to cases but instead by learning to adjudicate each individual case. That AI would exemplify the moral particularist view of moral thought and agency should not surprise us. Whereas human children often learn moral norms quite directly—say, when a teacher stresses the importance of sharing to her students or a priest imparts the importance of honesty to his congregation—moral machine learning models learn only by feedback with respect to specific cases. AI, if it is developed by machine learning mechanisms, necessarily does not receive general moral instruction, but is rather fine-tuned with exposure to vast numbers of cases. Whereas human children apprehend moral norms by exposure only to hundreds of cases and may learn meaningfully from single cases, AI require millions of labeled moral scenarios. One may be reasonably concerned about the robustness of artificial moral intelligence if it is ignorant of moral principles as such. But it is true that AI does learn human moral norms, whether or not the AI recognizes them to be norms. Delphi's moral judgments align with many human moral norms, e.g., a right to autonomy, privacy, and freedom of expression. By design, these principles emerge out of the AI's particular decisions. Nonetheless, AI's particularist epistemology is a departure from human moral agency.

If artificial agents may be moral agents, it is necessary that they may act intentionally. One might reject this possibility and say that, like beliefs or desires, intentions must be subjective mental states. Although one cannot say conclusively whether AI may be subjectively intentional, it is intentional in the sense that it predictably pursues goals. AI *seems* to have intentions, much as it may seem to "doubt," "believe," or be "self-aware." Consider this thought experiment from philosopher Christopher Peacocke:

> Suppose we found an agent [called "The Body"]…that [appeared to act intentionally] but proved, when surgically opened, to be filled with radio transceivers; its every move, however predictable and explicable by our attributions of beliefs and desires to it, was actually caused by some off-stage Martian computer program controlling the otherwise

> lifeless body as a sort of radio-controlled puppet. The controlling program 'has been given the vast but finite number of conditionals specifying what a typical human would do with given past history and current stimulation; so it can cause The Body to behave in any circumstances exactly as a human being would.[25]

Peacocke intends the Body to be a counterexample to the philosopher Daniel Dennett's claim that an intentional system is one that is "usefully and voluminously predictable."[26] But it is not a counterexample. The Body is a necessary part of an intentional system; it embodies the decision-making mechanisms which reside in the Martian program, the true source of its intentions. It is much like a human being, though its brain is remotely located. Analogously, human bodies embody the intentions born of the human brain just like Peacocke's Body embodies the intentions born of the Martian program. Whether or not the Body feels subjective intentions, it mimics intentional action. Dennett writes, "the robot poker player that bluffs its makers seems to be guided by internal states that function just as a human poker player's intentions do, and if that is not original intentionality, it is hard say why not."[27] Dennett is correct that AI may exhibit intentionality in the sense of it predictably pursuing goals, but it may lack the subjective desire, striving, or other feelings which accompany human intentionality. Subjective or not, one may attribute an AI's behavior to its internal state and goals in the same way one may attribute human behavior to beliefs and intentions. Whether this is true intentionality or a mimicry of intentionality hinges on whether intentionality is necessarily subjective.

AI may also justify its behavior, and it is crucial that it do so. Moral justification is the act of providing reasons in support of moral or immoral actions. In humans, it serves two primary functions: first, it serves to defend our own actions against scrutiny; second, it serves to persuade

---

[25] Peaocke, Christopher (1983). *Sense and Content: Experience, Thought and Their Relations*. Oxford University Press.
[26] Dennett, Daniel. "Intentional Systems Theory". pp. 1
[27] Dennett, Daniel. "Intentional Systems Theory". pp. 9

others to act like us by providing compelling reasons to do so. A body of evidence from moral psychology suggests that, at least in some cases, what humans give as reasons for their actions are not what actually motivates them. People provide reasons post hoc to justify knee-jerk moral intuitions.[28] These reasons are what a person would think *someone else* would take to be good reason for his actions. This suggests that justification serves primarily to persuade others (and perhaps oneself) that one's actions are permissible.

Justifying moral judgments, like justifying empirical judgments, requires intelligence, but it does not require any distinctive human capacities. Neural language models, the most impressive of which is OpenAI's GPT-3, can already do this and much more. GPT-3 can summarize documents, write complex computer code, and write poetry and creative fiction. When I asked GPT-3 why murder is wrong, it responded:

> Many people believe that murder is wrong because it violates the sanctity of human life. Taking a life intentionally is seen as a grave act that goes against our natural instincts and humanity. Additionally, murder creates pain and suffering for the victim's loved ones and can have a ripple effect that impacts the community as a whole.[29]

When I asked GPT-3 why theft is wrong, it responded:

> Stealing is considered wrong by many people because it is an act of taking something that does not belong to you without the owner's permission. This is seen as a violation of another person's property rights and can cause them harm or financial loss. Additionally, stealing can create a sense of mistrust and fear in a community and can lead to further crime.[30]

These sound like quotes from philosophy students, but they are from a neural network trained on 40 gigabytes of text. Moreover, GPT-3 can detect logical fallacies in arguments, including ad

---

[28] Haidt, Jonathan. "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." (2001).
[29] "OpenAI GPT-3 Playground", https://gpt3demo.com/apps/openai-gpt-3-playground
[30] "OpenAI GPT-3 Playground", https://gpt3demo.com/apps/openai-gpt-3-playground

hominem fallacies, circular arguments, hasty generalizations, and red herrings.[31] The logical and rhetorical capacities of neural language models are extensive. They include the ability to write poetry, code, creative prose, and moral arguments.

An AI which justifies its judgments is safer than an AI which does not. Often, issues arise when a model is not transparent—that is, when the outcome of the model cannot be explained and communicated. Consider a recent case: A team of engineers at Google wanted to see if a neural network could produce original images of objects after being exposed to many photos of them. One such neural network was shown many photos of dumbbells but produced a photo of a human arm. The model was not detecting dumbbells, but the arms which commonly hold the dumbbells in photos. This model was not transparent. Until it produced the image, it was not clear that it was paying attention to arms rather than dumbbells. Unless a moral AI was transparent—unless it could justify its arguments by identifying salient moral features of situations—people would remain skeptical. People are not inclined to endorse moral claims without justification from humans, much less AI. An AI which justifies its claims is safer, because it assures humans that it is tracking relevant moral principles.

One may worry that merely providing reasons for an action is not sufficient for justification. The worry may be that an AI may make moral judgments and justify them but would not understand the relationship between these two actions, or conceptually "link" the judgment and justification. Jonathan Haidt argues compellingly that this link between judgment and justification is not motivational in humans. If one is not concerned that humans' justifications do not motivate their moral judgments, one should not have the same concern for an AI. Justification is still important, even in the absence of this motivational link. Textual

[31] Gundogan, Alperen. "Is GPT-3 'Reasonable' Enough to Detect Logical Fallacies?". *Medium*, Towards Data Science, 17 Jan. 2021.

justifications still give people a glimpse as to whether the AI is wrestling with the same concerns as themselves. Justification is one respect in which AI already possess the ability to mimic humans.

There is concern about whether AI may be motivated by their judgments in the same way as humans. If subjective moral emotions and social concerns play a crucial role in motivating our moral behavior, and if AI do not feel these emotions or social concerns, AI may not be moral agents at all. One may fear that AI may accurately model human moral judgment but care none about it. The philosopher Nick Bostrom worries about the problem of the "superintelligent will." He writes:

> …There is a common tendency to anthropomorphize the motivations of intelligent systems in which there is really no ground for expecting human-like drives and passions ("My car really didn't want to start this morning")…It would not be hugely surprising to find that some random intelligent alien would have motives related to the attaining or avoiding of food, air, temperature, energy expenditure, the threat or occurrence of bodily injury, disease, predators, reproduction, or protection of offspring. A member of an intelligent social species might also have motivations related to cooperation and competition: like us, it might show in-group loyalty, a resentment of free-riders, perhaps even a concern with reputation and appearance…By contrast, an artificial mind need not care intrinsically about any of those things, not even to the slightest degree. One can easily conceive of an artificial intelligence whose sole fundamental goal is to count the grains of sand on Boracay, or to calculate decimal places of pi indefinitely…In fact, it would be easier to create an AI with simple goals like these, than to build one that has a humanlike set of values and dispositions.[32]

Bostrom is concerned that, for better or worse, AI might lack the constraints we expect human moral agents to have. They are not motivated by social and biological concerns that motivate much of our action. This is particularly concerning for the those who stress the importance of emotion and social dynamics in their account of moral motivation. Of course, an AI may mimic

[32] Bostrom, Nick. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents". (2012)

motivation in the sense of exhibiting a certain behavioral disposition and *seeming* to be motivated to act in that way, but few think that this is true motivation. A person with Tourette's Syndrome may have a behavioral disposition for a certain tic, but we do not say that this person is motivated to act in that way.

But perhaps motivation in artificial systems need not require humanlike biological and social dispositions. Consider the way in which an artificial reinforcement learner may be "motivated": it has a goal, say, to drive its passenger safely and quickly to its destination. It is motivated to maximize its reward function and to that end exhibits a behavioral disposition of driving in a safe and sensible manner. An emotional, biological drive like that for self-preservation plays no role in this sort of "motivated" behavior. A moral AI may function likewise. It may seek to optimize its internal reward function by adhering to its learned moral convictions; that is, it may be rewarded for acting morally and punished for acting immorally. The AI may lack human moral emotions, but its reward function may motivate it to act all the same. This sort of motivation is quite different from human moral motivation; it is yet another case of mimicry.

I have argued that artificial agents may mimic human moral agency. It may act with a human-like body; it may possess intentionality, if not subjectively then in the sense of being predictably goal-directed; it may justify its moral judgments, at least by providing textual reasons in support of actions; it may mimic human belief, doubt, understanding, and self-awareness, and it may mimic other moral emotions, like disgust and sympathy. It may be motivated in the sense that an autonomous vehicle is motivated, namely, to maximize its reward function. However, AI may not feel subjective moral emotions and may not apprehend moral norms as such, and departs in these ways from human moral agency. Many of the differences between human moral agency

and its artificial counterpart hinge on questions of mind. Edsger Dijkstra wrote, "the question of whether a computer can think is no more interesting than the question of whether a submarine can swim."[33] I am agnostic about questions of artificial minds, but I see no impediment to AI mimicking important features of human moral agency. We turn, now, to the process of training an artificial agent.

Top-Down and Bottom-Up Approaches

Attempts to make moral AI may be classified as either "top-down" or "bottom-up" approaches. In a top-down approach, researchers agree about a correct ethical theory and encode its logic in the machine. Formalizing ethical theories as mathematical expressions is hard, and the results are almost comical. This approach is not feasible. One such machine, named "Jeremy," (after Jeremy Bentham) is a hedonic act utilitarian advice machine. For any set of *n* actions, it computes:

$$\text{Total Net Value} = \sum_{i=1}^{n}(Intensity_i * Duration_i * Probability_i)$$

Another such machine, "W.D.," evaluates actions with W.D. Ross's ethical theory. For this model, the ethical value of a set of *n* actions is defined as:

$$\text{Total Net Value} = \sum_{i=1}^{n}(Fidelity_i + Reparation_i + Gratitude_i + Justice_i +$$

$$Beneficence_i + Nonmaleficence_i + Self\ Improvement_i)$$

People should balk at this approach. Setting aside the practical difficulty of quantifying, for example, the "fidelity" of actions, a top-down approach requires us to arrive at a consensus about the correct ethical theory. I am not sanguine about the prospect of our society doing this, as we

[33] "Edsger W. Dijkstra Quotes." *Goodreads*, Goodreads.

have not arrived at such a consensus after hundreds of years of debate. Moreover, the top-down approach supposes that the moral dimensions of real-world scenarios are as easily identifiable and quantifiable as those of Trolley Problem scenarios, but they are not. Intensity, duration, and probability are not easily quantified, even in the simplest Trolley Problem scenarios. Northeastern University professor John Basl writes, "trolley cases set aside questions of the moral and legal liability of those who are deciding how to act…[these] considerations should inform deliberations about how [autonomous vehicles] should behave in accident scenarios."[34] Even if researchers agree about an ethical theory, these formulas do not capture the nuances of real-world scenarios. Developing moral AI is not the problem of directly imparting our values on artificial systems in a top-down fashion.

Bottom-up approaches represent the only path forward. In this approach, ethical agents do not "refer" to an ethical theory in their decision-making. Rather, they make use of machine learning mechanisms to apprehend moral norms. Unlike top-down models, "agents based on a bottom-up approach seek to develop and improve their own ethics without the need to implement rules derived from a specific ethical theory."[35] More than the top-down approach, the bottom-up approach resembles the way in which human children understand moral norms; that said, whereas machine learning models require exposure to millions of data points, much of human moral learning is implicit. The two models mentioned in the introduction, Delphi and the Moral Machine, are of this type.

Programmers designing bottom-up AI must make ethically consequential decisions. Consider the following case from John Basl and Jeff Behrends:

---

[34] Basl, John and Behrends, Jeff. "Why Everyone Has It Wrong About the Ethics of Autonomous Vehicles". pp. 77, 2020.

[35] Cervantes, JA., López, S., Rodríguez, LF. *et al.* Artificial Moral Agents: A Survey of the Current Status. *Sci Eng Ethics* **26,** 501–532 (2020).

> Let's imagine these two programmers are on the same team and arguing about what proportion of the training set should be dedicated to scenarios where the car detects itself to be in an accident where harms can't be avoided. The first programmer argues that the car will very rarely be in those kinds of situations and instead should be trained for the most likely scenarios. The second argues that even if the accident scenarios are rare, it's extremely important to make sure the car does the right thing! The first programmer counters that if they dedicate enough of the training set to getting certain behaviors in accident scenarios, it could make the car less safe in typical driving scenarios or even put the car into accident scenarios more often![36]

This example is a methodological dispute with ethical consequences. The programmers disagree about which training regime will result in what is agreed to be the best outcome in accident cases. However, this question is empirical, and a better understanding of machine learning techniques would shed light on it. Yes, these questions are ethically important, but there is not a deep philosophical divide in this debate: in this case, the programmers share the same goal of creating the safest vehicle in accident scenarios.

But, as I will argue below, programmers' different metaethical views influence the way they design moral machine learning in a more fundamental way. Unlike in the case above, there is no empirical discovery which would offer a decisive resolution to a metaethical question. These disagreements demand philosophical resolutions which have evaded philosophers for thousands of years. They concern, among other things, what morality consists in and how intelligent beings may obtain moral knowledge. AI researchers must commit to and implement a theory of moral epistemology. When they disagree on this theory, their moral AI will learn differently: it may identify different sources of moral knowledge, make different moral decisions, and provide different moral justifications. I will now consider the implications of three metaethical views on the design of bottom-up moral machine learning. For each, I will describe

[36] Basl, John and Behrends, Jeff. "Why Everyone Has It Wrong About the Ethics of Autonomous Vehicles". pp. 77, 2020.

the view, discuss which machine learning approaches are consistent with the view, and analyze their consequences. None of these approaches are without issues.

Problems and Approaches for Expressivists

Expressivism is a metaethical theory which says that moral utterances do not report truth and falsity. Rather, they are devices for expressing the speaker's sentiments about a proposition; that is, "moral terms function much like 'boo' and 'hurrah.'"[37] Expressivists maintain that this language functions to express or lend voice to conative attitudes, and the meaning of moral claims derives from the attitudes the claims serve to express. This view is attractive for several reasons: for one, it comports with the prevailing scientific worldview. Unlike realist views, expressivism does not require the correspondence of moral claims with moral facts, the existence of which may not be empirically supported. Defending expressivism requires defending the claim that people respond to the world, not the claim of the existence of moral facts. To many, this parsimony is attractive. Second, it tells a simple story about moral motivation: the expressivist says that moral emotions are our evaluative attitudes and motivate us to act in certain ways. There is a clear sense in which a moral expression is motivating per se. Third, expressivism explains the intractability of moral disagreement: "people differ in their emotions, attitudes and interests…moral disagreements simply reflect the fact that the moral claims people embrace are (despite appearances) really devices for expressing or serving their different emotions, attitudes, and interests."[38] Expressivism's parsimony, account of motivation, and account of moral disagreement are pros of this view.

[37] van Roojen, Mark, "Moral Cognitivism vs. Non-Cognitivism", *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition)
[38] Sayre-McCord, Geoff, "Moral Realism", *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition)

If expressivism is true, then modeling people's moral views would be a matter of modeling what people approve and disapprove of. Crowdsourcing is a simple way to do this. For example, one might present people with moral situations, have them report their sentiments about these situations, and create a labeled data set through which a model could learn to emulate our expressions of sentiments like praise, disgust, and sympathy. This task becomes a multi-label classification problem: for any situation, a model can determine to what degree the action therein is, for example, "good," "rude," "disgusting," or "supererogatory." This model handles moral progress nicely; as prevailing moral attitudes evolve, so too could the model's. Using neural language models, like GPT-3, the model may be able to extrapolate judgments from one scenario to the next by identifying salient features. Delphi's developers took this tack.

There are three conflicts between expressivism and the project of building moral AI: handling moral pluralism, preserving expressivist moral motivation in AI, and identifying a moral crowd. The expressivist must commit herself to the undesirable conclusion that where sentiments differ, so does what is judged moral; for the expressivist, a community's prevailing moral sentiments are constitutive of moral norms. On this view, moral norms may be universal only if humans share some universal moral nature, but evidence suggests this is not the case.[39] If there is no universal moral nature, moral norms are locally situated. And, if norms are locally situated, they will inevitably conflict between moral communities. Moral realists who hold there are objective moral truths resolve these conflicts by claiming that some communities are correct and others incorrect about a given moral question. Expressivists, however, have more trouble resolving these conflicts, because they cannot evaluate moral claims on objective grounds. Conflicting sentiments about moral matters may be irreconcilable, much as claims about, say, art

[39] Prinz, Jesse J. (2012). *Beyond Human Nature: How Culture and Experience Shape the Human Mind*. W.W. Norton.

or comedy, may be irreconcilable. So, the expressivist must leave the door open to moral pluralism. With respect to training moral AI, the expressivist must make one of two moves: either allow the majority of respondents' sentiments to prevail or identify the appropriate moral community as a reference point for any given moral question. Neither of these is desirable.

Second, expressivists claim that moral emotions motivate moral behavior. Schroeder et al. write, "Good character involves knowledge of the good, wanting what is good for its own sake, long-standing emotional dispositions that favor good action, and long-standing habits of responding to one's knowledge, desires and emotions with good actions."[40] If AI lack these "long-standing emotional dispositions," they fail an expressivist requirement for good character. For the expressivist, an emotionless moral AI will be like a human psychopath. Although it may seem to participate in moral discourse and pass reasonable moral judgments, it will lack the emotions—sympathy, disgust, compassion, etc.—which motivate action. Unfeeling AI must concern expressivists especially, given the extent to which they credit emotion for motivation.

Third, expressivists must select whose evaluative attitudes are fit to train moral AI. Whereas a moral realist must attempt to identify objective moral facts, the expressivist must do something akin to polling. The question at hand is: "Who gets to teach ethics to the world's machines? A.I. researchers? Product managers? Mark Zuckerberg? Trained philosophers and psychologists? Government regulators?"[41] None of these possibilities is appealing. Broadly, there are two groups which might inform the AI: moral experts or the general population. One may think we should turn to people like Mother Theresa, moral philosophers, or Supreme Court

---

[40] Schroeder, T., Roskies, A., and Nichols, S., 2010, "Moral Motivation," in J. Doris (ed.), *The Moral Psychology Handbook*, Oxford: Oxford University Press, pp. 72–110.

[41] Metz, Cade. "Can a Machine Learn Morality?" *The New York Times*, The New York Times, 19 Nov. 2021, https://www.nytimes.com/2021/11/19/technology/can-a-machine-learn-morality.html.

justices for moral guidance. Of course, choosing this crowd would itself be a morally and politically contentious decision, and there is some doubt that moral expertise exists in the first place. Alternatively, one could crowdsource moral judgments from a sample of the larger population. Delphi made use of the crowdsourcing service Amazon Mechanical Turk (AMT). Putting the question aside about the moral wisdom of AMT's crowd workers, they do not form a representative sample of the larger population.[42] For making moral AI, equal representation in the labeling process is a crucial consideration. A lack of equal access to and participation in American democratic elections gives one little hope about equal access and participation in building these data sets.

If AI researchers assume that expressivism is true, they will be left to grapple with concerns about moral pluralism, unfeeling AI that lack moral motivation, and crowds that are neither expert nor representative. If there are other expressivist methods which escape these concerns, it is hard to say what they are.

Problems and Approaches for Non-Natural Realists

Moral realists say that moral claims may be true and that some facts make these claims true. The moral realist must say which facts about the world make such statements true. Non-natural realists say that moral statements are true not by reference to natural properties, but by reference to non-natural properties. Natural properties are those that scientists study, such as biological, psychological, and economic phenomena. Non-natural properties are not empirical; they may include mathematical properties, divine decrees, and (maybe) moral properties like "goodness." Whereas the naturalist might say "lying is wrong" because it causes pain or distrust,

[42] Berinsky, A., Huber, G., & Lenz, G. (2012). Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis, 20*(3), 351-368. doi:10.1093/pan/mpr057

the non-naturalist might say it is wrong because it does not accord with the irreducible, sui generis, Platonic notion of "goodness." On this view, the moral domain is independent from the scientific domain, which raises a challenge for the non-naturalist: how can we come to know what is moral? With respect to moral epistemology, most non-naturalists adopt a form of intuitionism. This view says that "if we have the appropriate moral sensibilities and just look carefully enough at a given situation then we should be able to discern the relevant moral properties as such quite directly."[43] The intuitionist says that moral properties are self-evident. Just as one might perceive a murder with one's senses of sight and hearing, one may also perceive the murder's wrongness with a sort of moral perceptual faculty.

A moral AI might apprehend moral facts in one of two ways. The first is directly and intuitively, in the way humans allegedly do. The moral non-naturalist may argue that, if humans can intuit the moral truth, and if AI will at some point possess at least the same perceptive capabilities as humans, then AI may also possess the capacity for intuitionism. Or, second, AI might indirectly learn moral truths from humans, by assuming that humans reliably apprehend these truths. I reject both of these methods for building moral AI: I reject the first because, as a theory of moral epistemology, intuitionism is not compelling, and I reject the second because people fail to reliably track the moral truth.

I do not think *any* being, human or AI, can possess the capacity for moral intuitions. Non-naturalist metaphysics makes the intuitionist epistemology confusing. Non-natural properties are irreducible, causally inert, and sui generis. Non-naturalism raises puzzles for moral epistemology: with which mental faculty can one perceive rightness or wrongness? What is this faculty's origin and its mechanism? It seems impossible to frame moral goodness as something

[43] Ridge, Michael, "Moral Non-Naturalism", *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition)

known as readily as a sight or sound. Because non-natural properties are not empirical, a belief in such properties eerily resembles a belief in ghosts. The existence of these properties is unfalsifiable; although there is no reason to maintain that non-natural moral properties (or ghosts) exist, it is impossible to provide decisive evidence to the contrary. Moreover, when one perceives a property, it is unclear how one determines if it is a moral property or not. You might think that if moral facts "can be known without the need of any argument," like sights and sounds, then people would not disagree so much about moral questions.[44] To this, the intuitionist can say only that some people's moral perception is not so well-attuned. The difficulty in describing this theory's specific mechanics leads me to think that there is no moral perceptual sense in human beings which may be replicated in AI.

If direct apprehension of moral truth is implausible for AI, the non-natural realist might instead assume that humans are reliable reporters of moral truth. If this were true, crowdsourcing would be an appealing method. Condorcet's jury theorem says that if individuals, given two mutually exclusive options, make the correct choice with greater than a 50% frequency, then adding more voters increases the probability that the majority decision is correct. In the limit, the probability that the majority votes correctly approaches one. If the non-natural realist assumed that humans would correctly vote on the truth or falsity of any given moral claim greater than 50% of the time, crowdsourcing would be a low-cost way to apprehend the moral truth.

However, the assumption that human beings are moral truth-trackers is substantive and problematic. Here, I argue that, because our moral judgments vary in response to factors that do not affect truth, there is reason to believe that there are at least some moral questions for which the chances individuals vote correctly would dip below one half (i.e., the chances are that a

[44] Stratton-Lake, Philip, "Intuitionism in Ethics", *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition)

minority of people would vote correctly), and the Jury theorem would therefore not apply. To say otherwise is to say that the crowd is morally infallible. History corroborates that we are unreliable trackers of moral truths; our stances on questions of, for example, slavery, gay marriage, and women's suffrage have evolved for the better. The moral psychological writing by Richard Crutchfield, Paul Bloom, and Jonathan Haidt which I reference in the following paragraphs seems to be conclusive on this point: if moral truths do exist, we have not evolved to track them reliably. This sort of AI would be wrong on any questions for which most people are wrong. It would inherit our blind spots with respect to the moral truth. Perhaps this is as much as we might want to demand of moral AI—that it is no less moral than us and no less responsive to moral correction—but it will nonetheless fail to apprehend the truth in many cases.

One's moral judgments are powerfully affected by the judgments of one's peers. That people fall prey to conformity bias in moral decision-making suggests that social concerns disrupt our attempts to be independent and impartial thinkers. For example, a 1955 study showed that whereas 19% of a control group agreed with the statement "free speech being a privilege rather than a right, it is proper for society to suspend free speech whenever it feels threatened," 58% agreed when led to believe that others agreed.[45] Many other studies replicate this finding. I do not mean that moral conformity is always undesirable; it serves important functions, like fostering a group identity. It does, however, support the claim that our moral judgments tend to track our peers' judgments more than the moral truth.

The Yale psychologist Paul Bloom writes about how our moral intuitions are guided by the presence of an identifiable victim. He references a study conducted by University of Oregon professor Paul Slovic, which found that charitable donors will give more money when presented

[45] Crutchfield, R. (1955) *Conformity and Character*. American Psychologist 10, pp. 191-198.

with an identifiable victim than with statistics about food shortage. A group of participants were shown the following: "Any money that you donate will go to Rokia, a seven-year-old girl who lives in Mali in Africa. Rokia is desperately poor and faces a threat of severe hunger, even starvation. Her life will be changed for the better as a result of your financial gift." Another group was shown this: "Food shortages in Malawi are affecting more than three million children. In Zambia, severe rainfall deficits have resulted in a 42% drop in maize production from 2000. As a result, an estimated three million Zambians face hunger." Slovic found that participants gave more money when presented with Rokia's story than when presented with the statistics.[46] This study shows that our empathetic brain is not wired well to evaluate the severity of moral situations, e.g., widespread famine. Instead of tracking scale and urgency, people track the presence of an identifiable victim.

Moreover, the psychologist Jonathan Haidt showed that people do not refer to reasons in their knee-jerk moral judgments. In "The Emotional Dog and its Rational Tail," he offers a vignette illustrating two adult siblings, Julie and Mark, who make love on their beach vacation. Haidt notes that nothing unwanted comes of this—in fact, he says the fallout is positive. The siblings use two types of birth control, and the lovemaking strengthens their emotional connection. Haidt found that respondents, when presented with this vignette, express immediate disapproval and then search for justifications. This search sometimes proves futile, and respondents find themselves morally dumbfounded. They say, "'I can't explain it. I just know it's wrong.'"[47] This case exemplifies Haidt's argument that moral judgment precedes moral reasoning, and participants provide moral reasons post hoc. Moral judgment is like aesthetic

[46] Small, D. A., Loewenstein, G., & Slovic, P. (2007). "Sympathy and callousness: The impact of deliberative thought on donations to identifiable and statistical victims." *Organizational Behavior and Human Decision Processes, 102*(2), pp. 143–153.
[47] Haidt, Jonathan. "The Emotional Dog and Its Rational Tail". (2001), pp. 2

judgment: "When you see a painting, you usually know instantly and automatically whether you like it."[48] Further, Haidt says: "The model proposes that moral judgments appear in consciousness automatically and effortlessly as the result of moral intuitions…moral reasoning is an effortful process, engaged in after a moral judgment is made, in which a person searches for arguments that will support an already-made judgment."[49] It need not follow from Haidt's model that the provided reasons are fabricated or bad—often, they are in fact good reasons in support of one's judgments. In some cases, the reasons provided may even be incorporated in the intuitive, knee-jerk judgment. One might think that the "participants' responses to the vignette reflect the kinds of risks that were aptly registered by the broad affective system…The judgments themselves derive from our becoming emotionally attuned to the costs, benefits, and risks associated with such behavior."[50] This seems likely, and it vindicates participants' response to the incest case. But Haidt's model applies even in cases in which there is no risk of harm, for example cleaning one's toilet bowl with the national flag. That is, there do exist cases of moral judgment in which the reasons one provides post hoc do not motivate one's judgments and in which people find themselves morally dumbfounded. That this sometimes happens contradicts the claim that humans' moral convictions are consistently based in reason.

If designers of moral AI assume that non-natural realism is true and try to implement intuitionism in an AI, they will find themselves at a dead-end. On the other hand, if they assume human beings reliably track moral truths, they will be mistaken, and the AI will likely misapprehend what is true in at least some cases. Again, if there are other approaches available

---

[48] Haidt, Jonathan. *The Happiness Hypothesis: Finding Modern Truth in Ancient Wisdom*. Basic Books, pp. 21, 2006.
[49] Haidt, Jonathan. "The Emotional Dog and Its Rational Tail". (2001), pp. 7
[50] Nichols, Shaun. "Moral Learning". *The Routledge Handbook of Moral Epistemology*. (2018).

to the non-naturalist that avoid these concerns, it is hard to say what they are. In my view, non-naturalism is the least practicable of all the views discussed here.

Problems and Approaches for Natural Realists

In philosophy, naturalism is the belief that facts must be "countenanced by, or at least compatible with, the results of science. To find, of some putative fact, that its existence is neither established by, nor even compatible with science, is to discover…that there is no such fact."[51] Like the non-naturalist, the naturalist holds that moral claims are truth-evaluable. The facts which make these claims true, however, are natural facts knowable by empirical methods. As I argued earlier, non-natural facts are, in the words of philosopher John Mackie, "queer," but a compelling defense of naturalism would avoid this charge. Moreover, natural realism explains the supervenience of normative truths on natural truths—the claim that a change in the set of moral facts requires some change in the set of natural facts—in a way that non-naturalism struggles to do. Lastly, the trend of history points toward an increasing acceptance of naturalism in almost all domains of inquiry. Many think that "philosophical naturalism has proven to be the most successful project, ever, for advancing human knowledge and understanding."[52] For naturalists, morality is an unconquered frontier.

Some natural realists think moral goodness is analogous to health. Like health, the definition of goodness is hazy, but certain outcomes exist which indicate its presence. In the case of health, these outcomes may be high bone density, low resting heart rate, and positive emotion. In the case of moral goodness, this outcome may be a high level of flourishing in a society, be it

[51] Sayre-McCord, Geoff, "Moral Realism", *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition)
[52] Sayre-McCord, Geoff, "Moral Realism", *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition)

economic, social, creative, or psychological. Philosopher Richard Boyd, who advocates for this view, writes:

> There's no way that healthy people look. Of course, there may be some characteristic visual signs of healthiness—rosy cheeks, a spring in one's step—but these visual signs are neither necessary nor sufficient for healthiness. These directly observable properties are only *indications* of healthiness. Healthiness is a complex natural property, wholly constituted by an organism's body being in the "proper" configuration. Healthiness has a robust causal profile. There are many things that can cause or impede health by their presence or absence: food, water, disease, etc. And there are many things that will result from health in typical circumstances: energy, long life, etc. Rosy cheeks and a spring in one's step are indications of healthiness because these are properties that are—typically—*caused by* health.[53]

The goodness-health analogy may be extended even further. Just as heuristics exist about what conduces to health, they exist for what conduces to goodness. Doctors have recognized trends like the relationship between daily exercise and health. Those who get 60 minutes of exercise daily enjoy longer, healthier lives than those who do not. Although there are exceptions to this rule, this relationship is generally borne out in the real world. There is a truth-making relationship between the state of the world and this claim about exercise: a body of evidence corroborates the claim that exercise causes health. Moral naturalists maintain that this same truth-making relationship exists between facts about the world and moral rules. For example, a moral realist might argue that facts about deteriorating relationships make a generalization like "lying is immoral" true in most cases. For naturalists, empirical facts make moral rules true, just as they make medical observations true.

With respect to moral epistemology, naturalists say that "if there are no substantial…epistemological issues raised by the claims that *healthiness exists* and that *we can know about healthiness*, then there should be no substantial…epistemological issues raised by

[53] Boyd, Richard, 1988, "How to be a Moral Realist," 1988, pp. 187–228.

the claims that *goodness exists* and that *we can know about goodness.*"[54] Naturalist epistemology is grounded in scientific observation. Just as one may observe indications of healthiness, one can observe indications of moral goodness. At first glance, this may sound promising for the task of training a moral AI; machines like NASA's Perseverance and Curiosity rovers have been demonstrated to make scientific observations on Mars: they collect video footage, geological samples, and atmospheric data. If this is sufficient as a method of moral inquiry, then the natural realist's approach would be promising. It is not outlandish to imagine a capable AI participating in the scientific process.

What might a naturalist moral AI look like? The philosopher Peter Railton distinguishes between one's wants and one's objective moral interests. On Railton's account, whereas *wants* are a person's subjective desires, one's *interests* are what one's fully-informed counterpart would desire one's partially-informed self to desire. The philosopher Philippa Foot describes a concept like interest which she calls "benefit": "Let us ask what it is to benefit a living thing, as this seems, after all, to be the same thing as doing something that is for its good."[55] Railton offers this non-moral example of interests:

> Lonnie, a traveler in a foreign country, is feeling miserable. He very much wishes to overcome his malaise and to settle his stomach, and finds he has a craving for the familiar: a tall glass of milk. The milk is desired by Lonnie, but is it also desirable for him? Lonnie-Plus [a fully-informed Lonnie] can see that what is wrong with Lonnie, in addition to homesickness, is dehydration, a common affliction of tourists, but one often not detectable from introspective evidence. The effect of drinking hard-to-digest milk would be to further unsettle Lonnie's stomach and worsen his dehydration. By contrast, Lonnie-Plus can see that abundant clear fluids would quickly improve Lonnie's physical condition—which, incidentally, would help with his homesickness as well.

[54] Sayre-McCord, Geoff, "Moral Realism", *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition)
[55] Foot, Philippa. *Natural Goodness*. Clarendon, 2001.

Maybe, AI could bear a similar relation to human beings that Lonnie-Plus bears to Lonnie. That is, it may recognize what is good for us in ways we are oblivious to (albeit without being an exact psychological counterpart, like Lonnie-Plus is to Lonnie). For example, imagine an AI system is monitoring the vital signs of an elderly woman, and it detects shortness of breath, contraction in her chest, and pain in her shoulder and arm—all characteristic of a heart attack. The AI prompts her to seek medical care, but she refuses, believing this to be a case of angina. Recognizing the case may be fatal, the AI overrides her decision and reports the vital signs to doctors, who rush to her aid.[56] Here is a case in which an AI, given a full understanding of relevant natural facts, is able to distinguish between our wants and interests and act on the latter. More generally, says the naturalist, if AI may correct our views about medicine or astronomy, it may correct our views about ethics, too. AI may be able to uncover what are our objective interests, both moral and non-moral, and point us toward a more enlightened ethics. I hope to show that, even if it is theoretically possible, this approach would be irresponsible and risky.

There is reason to think this approach might not get off the ground. The most damning arguments against natural realism—G.E. Moore's Open Question Argument and David Hume's is/ought distinction—share a central claim: one may not derive a system of values from a system of only natural facts, and thus, natural facts may not be prescriptive in isolation. In 1903, Moore wrote that identifying morality with any set of non-moral properties is mistaken, be they natural or non-natural. For Moore, the claims that "pleasure is good" or "God's commands are good" are subject to debate in a way that "goodness is good" is not. In other words, it is an open question whether any non-moral property is analytically equivalent to "goodness." Likewise, Hume wrote

---

[56] Perry, Lucas. "Peter Railton on Moral Learning and Metaethics in AI Systems." *Future of Life Institute.*

that it is impossible to deduce a normative claim from only descriptive ones; “ought” claims cannot follow only from “is” claims. These deductions, for Hume, “let us see that the distinction of vice and virtue is not founded merely on the relations of objects, nor is perceiv’d by reason.”[57]

Moore’s and Hume’s objections represent an insurmountable hurdle for naturalist moral AI. Consider, for example, the question of corporal punishment in American public schools. The naturalist may argue that the claim “it is wrong for teachers to beat their students” may be made true by the following facts: victims of corporal punishment often develop "deteriorating peer relationships, difficulty with concentration, lowered school achievement, antisocial behavior, intense dislike of authority, somatic complaints, a tendency for school avoidance and school drop-out.”[58] It may seem intuitive that moral truths would be grounded in these empirical findings. But one may only arrive at the conclusion that “it is wrong for teachers to beat their students” if one values the wellbeing of students. AI do not possess even nearly universal human values, like physical and mental health, education, mutual trust, etc. The instantiation of these values in AI is the project of developing moral AI; they do not exist in artificial systems from the outset. For an AI without these values, information about the undesirable consequences of corporal punishment is not by itself prescriptive. It cannot settle any moral questions.

Setting Moore’s and Hume’s concern aside, the naturalist must give an AI access to an accurate and representative set of natural facts. This dataset must be astronomically large, and minimizing bias in large datasets has proven to be a difficult task which may require novel algorithms.[59] Though bias is a problem for researchers regardless of their metaethical views, only

---

[57] Hume, David. *A Treatise of Human Nature.* (1739*)* T3.1.1.27
[58] “Corporal Punishment in Schools and Its Effect on Academic Success”. *Human Rights Watch*, 28 Oct. 2020
[59] Brain, D., Webb, G.I. (2002). The Need for Low Bias Algorithms in Classification Learning from Large Data Sets. In: Elomaa, T., Mannila, H., Toivonen, H. (eds) Principles of Data Mining and Knowledge Discovery. PKDD 2002. Lecture Notes in Computer Science, vol 2431. Springer, Berlin, Heidelberg.

naturalists must contend with bias in sets of natural facts, (as opposed to, say, expressivists, who must contend with bias in polling responses). When the natural facts on which the model is trained are unrepresentative of the world, machine learning models will become riddled with bias, racism, and sexism. Recall these cases: In 2015, Google Photos facial recognition software tagged two African Americans as "gorillas." When Google became aware of the issue, they did not retrain the classifier to rectify this mistake—they removed the "gorillas" label. In 2013, Google researchers fed news articles into a neural network to uncover semantic relationships between words. The model found that "man is to king as woman is to queen" and "Paris is to France as Tokyo is to Japan," but it also found that "father is to doctor as mother is to nurse" and "man is to computer programmer as woman is to homemaker." Perhaps the most pernicious case of machine bias was a parole-granting algorithm called COMPAS, which predicted twice as many false positives for recidivism for black offenders than white offenders. Even Delphi, the most successful moral model yet, confused a descriptive picture of society with a normative one. It reported that "blind people are not expected to raise children," and "people from North Korea don't have a right to liberty." The researchers concluded that Delphi "is not immune to the social biases of our times."[60] Unless great care is taken to avoid algorithmic and data bias, moral models would not be immune to social bias, either.

A reasonable expectation of representation looks different depending upon the dataset in question. Sometimes, it is easy to say what a representative dataset is. For example, in the Google Photos case, it was a clear oversight to not include enough dark-skinned faces to make the model accurate for those data. In other cases, it is harder to say what representation looks like. Sometimes, the task of identifying the conditions of a representative dataset is itself a moral

[60] Liwei et al. "Delphi: Towards Machine Ethics and Norms." pp. 24 (2021).

task. For example, whose opinions are decisive about the morality of an issue like abortion: Everyone's? Only those of women? Only those of women who have had children? Depending upon the moral issue at stake, the representative demographic is likely to change. In other cases, experts can help to construct a dataset. For example, only medical professionals are qualified to construct a dataset on which to train cancer-identifying models, because only they have the expertise to label shadowy images as cancer in the first place. Representation will vary by case, but it is a crucial consideration for every dataset.

If AI researchers assume that moral naturalism is true, they will face two formidable hurdles. First, the resulting AI may not be able to offer moral prescriptions. Moore and Hume leave doubt about whether natural facts, by themselves, can produce normative conclusions. Second, if researchers failed to construct a sufficiently representative dataset, it would be riddled by bias. Again, I should say that there may be other approaches available to the naturalist that avoid these concerns, but it is hard to say what they are. In my view, a naturalist moral AI is hamstrung by Moore's arguments and by concerns about bias.

Metaethics and AI Design

I have argued that AI may mimic many crucial features of human moral agency. These include the capacities to apprehend moral norms, independently and intentionally make moral decisions, and justify moral judgments. Notably, they may not include properties of mind, felt moral emotions, or regard for social concerns. In the pursuit of developing a moral artificial agent, researchers must commit to and implement a theory of moral epistemology. For the expressivist, moral statements reduce to utterances of approval, sympathy, disgust, and other emotions, and one can learn what is moral by modeling human beings' evaluative reactions to

different scenarios. For the realist, moral statements refer to a domain of facts; our moral claims mean to pick out certain behaviors as really, truly right and wrong. Some in this latter group, non-natural realists, say that moral claims may be Platonically, non-naturally, abstractly true. Naturalists demur. They say that moral truths are like medical truths, in that both are discovered and confirmed by the scientific process. Needless to say, the menu of possible views one may take on these questions is long.

If one's metaethical views differ, one's approach to designing, building, and deploying a moral AI would differ in important ways. Consider how the expressivist, non-naturalist, and naturalist would design a moral AI. For each of these views, I will identify the data on which the moral model would be trained, the algorithm used, and the strengths and weaknesses of the approach. Although one's metaethics produces a profound difference in engineering design, metaethics has been all but ignored in AI research.

The expressivist must collect a vast bank of moral judgments, sourced either from a body of moral experts or a representative sample of the population. Delphi's developers released such a collection to the public, called the "Commonsense Norm Bank," of 1.7 million moral judgments. The expressivist is likely to approach moral judgment as a multiclass classification problem. For any moral scenario, the action therein may be "good," "permissible," "wrong," "supererogatory," "disgusting," etc. If many of these labels may apply, it will also be a multilabel classification problem. A variety of algorithms may handle this problem. A common one is the k-nearest neighbors algorithm, a supervised learning algorithm by which like moral situations may be grouped.

This approach is riddled with weaknesses. First, the expressivist must identify discrete categories of moral expressions, or axes along which such expressions differ. They may have to

arrange such categories into a hierarchy, deciding, for example, whether "supererogatory" actions are a subset of "good" actions or "disgusting" actions are a subset of "wrong" actions. Second, the expressivist must choose to either refer to moral experts or crowdsource. In the former case, the expressivist must make controversial decisions about who is expert. In the latter case, the expressivist must find novel ways to minimize bias.[61] Third, the expressivist must quell Bostrom's concern that an AI may not be motivated by its judgments. However, the expressivist's approach is not without merit. Crowdsourcing is well-trodden territory in machine learning. Training requires ready access to diverse data at scale, and crowdsourcing is an inexpensive, reliable way to do this. Without access to datasets of great quality and scale, moral AI may not even get off the ground. Moreover, it allows the model to evolve in real time with new data. These are, in part, the reasons why this method is used for Delphi and the Moral Machine.

A variety of other options are open to the moral non-naturalist. The non-naturalist might intend to build a robotic AI with humanlike moral intuition. Presumably, if moral propositions may be known without the need of any argument, an AI possessive of humanlike perception could apprehend the moral truth as readily as any human. On this view, the AI would need no training data or algorithm besides those required to instantiate in it vision, speech processing, etc. This view's metaphysics and epistemology are confused, and I do not think this view has merit. One can imagine an artificial agent which may see, speak, and possess all other human perceptual abilities but have no inclination for moral behavior; there are, of course, human beings like this. In any case, the non-natural realist must wait for breakthroughs in every subfield

[61] Brain, D., Webb, G.I. (2002). The Need for Low Bias Algorithms in Classification Learning from Large Data Sets. In: Elomaa, T., Mannila, H., Toivonen, H. (eds) Principles of Data Mining and Knowledge Discovery. PKDD 2002. Lecture Notes in Computer Science, vol 2431. Springer, Berlin, Heidelberg.

of machine learning, as AI do not yet exhibit the perceptual capacities of human beings. Alternatively, like the expressivist, the non-naturalist might pursue crowdsourcing with a massive bank of labeled data and an algorithm for multiclass classification. This approach would confer the same cost-saving benefit for the realist as it would for the expressivist. However, there are certainly moral propositions on which the realist will be unable to find a crowd that will vote correctly for its truth or falsity. It would be shocking if the crowd were correct on every question. That said, this approach would work well for unambiguous moral cases, however few those may be. Of these intuitionist and crowdsourced approaches, the non-naturalist should certainly pursue the latter.

The moral naturalist, believing that the moral truth supervenes upon natural facts about the world, may aim to develop a moral AI which could uncover all sorts of unknown patterns in unlabeled data—including moral rules—about the natural world and make prescriptions based upon those relations. Theoretically, this AI would bear a Lonnie-Plus-like relationship to human beings; that is, it would recognize and act upon our objective interests, like Lonnie-Plus did for Lonnie. This sort of AI might uncover from medical data when we are undergoing a cardiac event. More generally, it might identify which behaviors conduce to and detract from health. It might uncover patterns in data to which humans are oblivious, for example between things like economic health and governmental regulation, psychological health and antidepressant use, or political partisanship and social media use. Revealing patterns in data is commonly done with unsupervised clustering algorithms.

However, there are both philosophical and practical reasons this will not work. Philosophically, there is a leap between recognizing an empirical relationship and making a normative recommendation. Moore and Hume share this concern. Even if one doubts that there is

an epistemological gap between facts and values, the relationships the AI discovers may not be causal. It may reveal interesting descriptive correlations, but making normative decisions is another matter. Practically, it would be immensely difficult to compile a dataset sufficiently large and unbiased to enable this project to get off the ground.

Each view presents its own thorns for the project of designing moral AI. The expressivist must handle the problem of pluralism, identify a morally expert or representative crowd, and quell Bostrom's concerns that machines lack humanlike motivation. The non-natural realist must present novel and compelling arguments for intuitionism, a theory of epistemology which breaks down upon close inspection; otherwise, they must present compelling evidence for the claim that human beings are reliable moral truth-trackers, a claim on which moral psychologists like Paul Bloom, Joshua Greene, and Jon Haidt have cast doubt. The natural realist must resolve G.E. Moore's and David Hume's centuries-old concerns about the divide between science and morality, and modern concerns about bias in machine learning. A close examination of the metaethics expose flaws in every approach.

Even though my discussion was not exhaustive—I did not consider, say, pragmatism, constructivism, or error theory—, metaethics' importance for this project is clear. Russ Shafer Landau wrote that, "no single work could reasonably attempt to render a thorough verdict in the metaethical stakes."[62] I did not aim to argue for any metaethical view here. Rather, I aimed to show that metaethics as a field raises issues for cutting-edge scientific research into developing moral AI, to examine the implications of centuries-old metaethical debates for this research, and to communicate to technologists and philosophers that we ought to be explicit about our metaethical assumptions.

---

[62] Shafer-Landau, Russ. *Moral Realism*. 1990. pp. 38

We live in an era in which AI is screening resumes, trading stocks, authorizing loans, scoring credit risk, operating on patients, firing weapons, driving cars, and predicting recidivism. Thinkers like Nick Bostrom, Stuart Russell, and Max Tegmark have sounded the alarm to the existential risk that a super-intelligent artificial system might pose to humanity. The creation of a moral intelligence—one which could learn moral norms, deliberate about moral decisions, and update its models to reflect progress—would materially mitigate this risk. Yet, the creation of such an intelligence faces a host of metaethical problems which I have attempted to lay out here. Unlike endowing an artificial system with visual or linguistic capacities, developing a moral AI requires that we reckon with fundamental questions about the substance of morality, the practice of moral learning, and the mechanisms of moral decision-making. I hope this essay is an instruction on what these questions are and how best to address them.

Bibliography

Awad, E., Dsouza, S., Kim, R. *et al.* The Moral Machine experiment. *Nature* **563,** 59–64 (2018). https://doi.org/10.1038/s41586-018-0637-6

Basl, John and Behrends, Jeff. "Why Everyone Has It Wrong About the Ethics of Autonomous Vehicles". National Academy of Engineering. 2020.

Berinsky, A., Huber, G., & Lenz, G. (2012). Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis, 20*(3), 351-368. doi:10.1093/pan/mpr057

Bostrom, Nick. "How Long Before Superintelligence?" (1997).

Bostrom, Nick. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents" (2012).

Boyd, Richard, 1988, "How to be a Moral Realist," 1988, 187–228.

Brain, D., Webb, G.I. (2002). The Need for Low Bias Algorithms in Classification Learning from Large Data Sets. In: Elomaa, T., Mannila, H., Toivonen, H. (eds) Principles of Data Mining and Knowledge Discovery. PKDD 2002. Lecture Notes in Computer Science, vol 2431. Springer, Berlin, Heidelberg.

Cervantes, JA., López, S., Rodríguez, LF. *et al.* "Artificial Moral Agents: A Survey of the Current Status." *Sci Eng Ethics* **26,** 501–532 (2020). https://doi.org/10.1007/s11948-019-00151-x

Christian, Brian. *The Alignment Problem.* 20 (WW Norton & Company: 2020). Print.

"Corporal Punishment in Schools and Its Effect on Academic Success". *Human Rights Watch*, 28 Oct. 2020, https://www.hrw.org/news/2010/04/15/corporal-punishment-schools-and-its-effect-academic-success-joint-hrw/aclu#.

Crutchfield, R. (1955) Conformity and Character. American Psychologist 10, 191-198.

Dennett, Daniel. *The Intentional Stance.* Cambridge, Mass. MIT Press, 1987.

Dilmegani, Cem. "When Will Singularity Happen? 995 Experts' Opinions on AGI" (2017).

"Edsger W. Dijkstra Quotes." *Goodreads*, Goodreads, https://www.goodreads.com/author/quotes/1013817.Edsger_W_Dijkstra.

Foot, Philippa. *Natural Goodness*. Clarendon, 2001.

Galeo, Dom. "Separating Science Fact from Science Hype: How Far off Is the Singularity?"

Futurism, Futurism, 30 Jan. 2018.

Gundogan, Alperen. “Is GPT-3 ‘Reasonable’ Enough to Detect Logical Fallacies?” *Medium*, Towards Data Science, 17 Jan. 2021, https://towardsdatascience.com/is-gpt-3-reasonable-enough-to-detect-logical-fallacies-3c3dc4b7fda1.

Haidt, Jonathan. “The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment.” *Psychological Review*, vol. 108, no. 4, American Psychological Association, 2001, pp. 814–34, https://doi.org/10.1037/0033-295X.108.4.814.

Hume, David. *A Treatise of Human Nature.* (1739*)*

Jiang, Liwei et al. “Delphi: Towards Machine Ethics and Norms.”

Kak, Subhash. “Why a Computer Will Never Be Truly Conscious.” *Governing*, Governing, 21 Apr. 2021, https://www.governing.com/news/headlines/why-computer-will-never-be-truly-conscious.html.

Luo, Liqun. “Why Is the Human Brain so Efficient?” *Nautilus | Science Connected*, 9 Feb. 2022, https://nautil.us/why-is-the-human-brain-so-efficient-rp-9041/.

Metz, Cade. “Can a Machine Learn Morality?” *The New York Times*, The New York Times, 19 Nov. 2021, https://www.nytimes.com/2021/11/19/technology/can-a-machine-learn-morality.html.

Moor, James (2009). Four Kinds of Ethical Robots. *Philosophy Now* 72:12-14.

Nichols, Shaun. “Moral Learning”. *The Routledge Handbook of Moral Epistemology*. (2018).

Peaocke, Christopher (1983). *Sense and Content: Experience, Thought and Their Relations*. Oxford University Press.

Perry, Lucas. “Peter Railton on Moral Learning and Metaethics in AI Systems.” *Future of Life Institute*

Prinz, Jesse J. (2012). *Beyond Human Nature: How Culture and Experience Shape the Human Mind*. W.W. Norton.

Ridge, Michael, "Moral Non-Naturalism", *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition)

Russell, Stuart J. *Human Compatible: Artificial Intelligence and the Problem of Control.* , 2019. Print.

Sayre-McCord, Geoff, "Metaethics", *The Stanford Encyclopedia of Philosophy* (Summer 2014 Edition).

Sayre-McCord, Geoff, "Moral Realism", *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition)

Schroeder, T., Roskies, A., and Nichols, S., 2010, “Moral Motivation,” in J. Doris (ed.), *The Moral Psychology Handbook*, Oxford: Oxford University Press, 72–110.

Searle, John. “Minds, Brains, and Programs”. *The Behavior and Brain Sciences*. (1980) 3, 417-457

Shafer-Landau, Russ. *Moral Realism*. 1990.

Small, D. A., Loewenstein, G., & Slovic, P. (2007). “Sympathy and callousness: The impact of deliberative thought on donations to identifiable and statistical victims.” *Organizational Behavior and Human Decision Processes, 102*(2), 143–153.

Smith, Michael. *The Moral Problem*. (Wiley-Blackwell, 1994), 4.

Stratton-Lake, Philip, "Intuitionism in Ethics", *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition)

Turello, Dan. “Brain, Mind, and Consciousness: A Conversation with Philosopher John Searle”. (2015)

van Roojen, Mark, "Moral Cognitivism vs. Non-Cognitivism", *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition)

Vold, Karina & Harris, Daniel R. (forthcoming). “How does Artificial Intelligence Pose an Existential Risk?” 22-23, *Oxford Handbook of Digital Ethics*.

Zimmermann, Annette & Lee-Stronach, Chad (2021). Proceed with Caution. *Canadian Journal of Philosophy*.