AI Ethics: The Thin Line Between Computer Simulation and Deception

Larry J. Crockett
Augsburg University, Minneapolis MN, USA
crockett@augsburg.edu
DOI: 10.34190/ECIAIR.19.032

Abstract: Elon Musk famously said we are "summoning the demon" by developing AI but the rest of us do not agree what the demon is. It might be the threat of a malevolent superintelligence or the immoral enslaving of artificially intelligent agents. Against AI, the Machiavellian Intelligence Hypothesis argues it is social competition that has driven primate intelligence so that ethics originates in the advantages offered by lying better than other liars. We should not be surprised, as a result, by the "replicability crisis" in science, stemming partly from methodological difficulty but often from fraud. The "new moral science" attempts to build an ethics freed from conflicting human intuitions and buttressed by the reliability of a science now undermined by this crisis. Compounding the problem, complex adaptive systems cannot be captured by theories or algorithms simpler than the systems, with the result that both understanding and predictability suffer. Coarse-grained simulations of such systems restore some predictability but at the price of inviting misunderstanding and fraud. Just as conventional ethics arose in deceptive behavior such as lying, so AI ethics originates in the treacherous relationship between close and coarse graining in simulation so widely used in artificial intelligence. Musk's demon is closer at hand than a destructive superintelligence or smart robots we unwittingly enslave. Instead it lies in the deeply mistaken way we currently train people in technologies such as artificial intelligence. It is therefore much more urgent.

Keywords: new moral science, Machiavellian Intelligence Hypothesis, intuition, replicability crisis, coarse graining

1. Introduction

The headline on *Business Insider*'s front page is not subtle: "'Shame on Mark Zuckerberg': Facebook enraged lawmakers again after evading questions about its year from hell" (Hamilton 2019), but the headline understates the problem. Technology's ethical deceptions range across the industry and over longer spans of time, especially in terms of its use of AI and big data. It is unlikely the ethics "hell" will last only a year.

How did we get to this deplorable juncture? Artificial intelligence has been pursued at least since the 1950s when John McCarthy coined its provocative name, but most of the decades since that time AI has not been as publicly controversial. In fact, any number of times AI was judged to be a failure (Dreyfus 1972, 1992), derailed by a misunderstanding of intelligence (Crockett 1994). A philosopher, Dreyfus taught at MIT and encountered early AI efforts. In an interview (Grimes 2017), Dreyfus said, "They said they could program computers to be intelligent like people … they came to my course and said, 'We don't need Plato and Kant and Descartes anymore. …we're empirical [and] we're going to actually do it.'"

Classical challenges to whether AI will become genuinely intelligent have faded in favor of debates over the ethical consequences of when they do. Notable examples include Elon Musk who said, "With artificial intelligence we are summoning the demon" (McFarland 2014) and Stephen Hawking who warned "it could be the worst thing that has ever happened to humanity" (Osborne 2017). Oxford's Nick Bostrom's *SuperIntelligence: Paths, Dangers and Strategies* (2014) stands as the foremost book-length warning of what will happen when "superintelligent AI" is achieved, largely assuming it is inevitable.

This paper will argue we face a different, more immediate danger—that we will prematurely cede a wide range of ethical judgments to AI when the evidence and the argument favor the conclusion that ethical judgments are better left to people well schooled in a wide variety of scientific and ethical disciplines. The immediate challenge, therefore, is a sweeping revision of our technological education.

2. Science and the New Moral Science

The European Enlightenment (1650-1800) represented a sea change in how scholars believed we should reach beliefs about the world and behave in human communities. Appeals to method, notably scientific method, replaced appeals to authorities, such as classical texts. As chronicled by Hunter and Nedelisky (2018, 47), teleological Aristotelian ideals were replaced with a "comprehensive science of man and nature"—and with that, the goal of a "scientific morality."

Perhaps the most important ingredient in this science of man and morality was the philosophical doctrine of *naturalism*, the view that everything that exists is best understood in scientific terms. It is important to note that naturalism is not a view derived *from* science but instead is applied *to* it. Such a view cannot be derived from data or scientific theory alone, which is what makes it philosophical.

Naturalism competes with traditional views that there are other nonscientific, non-empirical bases for truth, knowledge, understanding, and wisdom. These include intuition, common sense, introspection, cultural traditions, and, notably, pure reason. The pivotal question is whether human intuition is essential to the practice of science, mathematics and computer science, and ethics. Paralleling the history of mathematics, in fact, ethics since the Enlightenment has often been dominated by the effort to eliminate intuition.

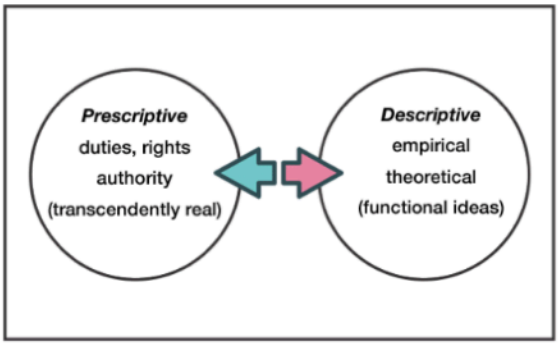

Figure 1: Bentham's Felicitous Calculus and the Science of Morality

Bentham's "felicitous calculus," illustrated in Figure 1, likely inspired by Leibniz's earlier view that pleasure and pain can be quantified and calculated, held that the value of a given pleasure or pain is measurable in terms of intensity, duration and proximity. The morally *prescriptive* is derivable from the empirically *descriptive* without intuition. Leibniz, a diplomat mathematician-philosopher in a time of war, argued humanity needs a system of symbols that encompasses the full scope of human thought, which can compute answers to all questions so that all will agree with the results. Ethics, for Leibniz, is therefore both scientific and computable.

The question for these early moral scientists was: Can science do for morality what it did for chemistry and physics—specifically, resolve conflicting views with empirical evidence and theories? But like an unnoticed Trojan horse, David Hume's famous Law was seeded into the mix: "No ought from is." No moral conclusions can be derived from factual premises alone. In Hume's view, there is no objective good or evil outside our own feelings. As a result, an unstable synthesis emerged in the moral scientists who pre-dated the modern era: Hume's mind-focused sentimentalism, Darwin's evolutionary account of humanity, and Bentham's utilitarianism shaped a naturalism committed to the empirical study of ethics.

2.1 Natural Science and the Naturalistic Fallacy

But this unstable synthesis was shaken at the turn of the 20th century. Moore (1903) argued that reducing ethical terms to properties of the natural world, namely, those studied by natural science, commits the *naturalistic fallacy*. For instance, to take the meaning of the word "good" to be definable, explanatorily reducible, to the terms used in natural science commits this fallacy since something essential is inevitably lost. Evolutionists, to sharpen the example, often defined "good" in terms such as "highly evolved." But any claim that "good" (moral concept) can mean something like "highly evolved" (biological concept) will always generate what Moore called "the open question." For example, "But is it good to be highly evolved?" That is, the naturalistic reduction always engenders an open question, which means the translation fails. In other words, there are no natural terms that can be substituted for moral terms without moral loss.

For the better part of a century after Moore wrote, it was widely accepted by ethicists that ethical properties are not replaceable with natural properties (Hunter and Nedelisky, 2018, 71). That is, "good" is irreducible ethically, it exerts a non-replaceable constraint of other ethical concepts. As a result, the Enlightenment-inspired idea that ethics should become a science of morality waned—but only for a time.

2.2 The New Moral Science, Gödel Incompleteness and the Return of Intuition

The "new moral scientists" (Hunter and Nedelisky 2018) see the role of science as discovering how moral psychology and brain chemistry can both explain and shape moral ideas and behavior. Advances in biology, cognitive science and AI herald a breakthrough in ethics as well. Ignoring Moore, moral judgments should be based on empirical evidence and scientific theory rather than intuition. In other words, "moral intuitions are permitted only if a place can be found for them within a naturalistic metaphysic" (Hunter and Nedelisky 2018, 170). A prominent example of the new moral science is Pinker (2018).

Given the unprecedented surge of interest in AI, the question naturally arises regarding the relationship of the new moral science to AI. Turing's (1936) path-breaking understanding of algorithmic processes and with it the potential to make a single "all-purpose" machine, originated in his attempt to answer Hilbert's famed *Entscheidungsproblem*, namely, the problem of whether a single machine can determine the truth or falsity of arbitrary mathematical questions. In parallel with the later moral scientists, Hilbert saw the origin of modern mathematical paradoxes to be math's reliance upon intuition. The foundation of math should be purely formal, with no appeal to intuition.

So there is an important parallel between Hilbert's attempt to squeeze intuition out of mathematical proof and the attempt to build ethics scientifically by eliminating appeals to non-empirical ideas such as our intuitive sense of good. But Hilbert's program was spectacularly derailed by Gödel's incompleteness theorems (Gödel 1931). Russell and Norvig (2015, 357) concur that Gödel incompleteness has been widely debated in AI. Gödel argued that mathematics inevitably involves intuitive description of an unchanging mathematical realm of pure ideas. Note the parallel between Moore and Gödel. "Truth" is not reducible to provability for Gödel just as "good" is not reducible to scientific terms for Moore. Intuition is critical to both.

2.3 The New Moral Science and Artificial Intelligence

Gunkel (2012) may provide the most comprehensive treatment of AI ethics. He skillfully argues that moral theory has been indefensibly anthropocentric, resisting claims, for example, that animals and machines might warrant recognition as moral agents. Moreover, Gunkel endorses Hunter and Nedelisky's (2018) characterization of modern moral science since ethical theory is not the ongoing discovery of eternal Platonic forms but the historical product of empirical investigation and specific times and places, reflecting the self-correcting characteristic of any science.

But Gunkel's claim, that the principal AI ethical problem involves enslaving morally aware machines, is unjustified. Bringsjord and Govindarajulu (2018) argue adroitly that we are nowhere near Turing Test intelligence (or Bostrom's superintelligence). They remind us of the long and deep relationship of philosophy to questions associated with what we call "artificial intelligence": Aristotle's syllogistic theory; Descartes' (1637) anticipation of the Turing Test; Turing's (1950) celebrated *Mind* paper; and Searle's (1980) widely cited "Chinese Room Experiment," all bear on both McCarthy-era AI and more recent neural-net AI. Descartes trenchantly observed, "But it never happens that it arranges its speech … as even the lowest type of man can do." Turing incorrectly predicted we would be satisfied by 2000 that machines are intelligent and we could ask the philosophers to step aside. Bringsjord and Govindarajulu (2018) underscore that Descartes "is certainly carrying the day" since "AI simply hasn't managed to create general intelligence" and cannot manage even a credible, involved conversation with a child.

Hence, the philosophical questions remain. Figure 2 illustrates Hunter and Nedelisky's (2018) philosophical rebuttal of the new moral science. The levels in Column 1 usefully correspond to Wisdom, Knowledge and Information (Crockett 2002). Level 1 contains specific moral concepts of the good and what should be done; Level 2 yields evidence for or against some moral claim or theory; Level 3 contains scientific descriptions of the origins of morality (for example in the brain) or how moral judgment is purportedly embodied in brain neurons.

The distinction between prescriptive and descriptive in Column 2 reprises Hume's resilient distinction between ought and is. The gulf between the prescriptive and descriptive remains unbridgeable. Last, Column 3 illustrates the claimed move from computation as information manipulation in Level 3, to AI in Level 2, to the profound question of the possible moral agency of AI in Level 1. The new moral science insists movement up the levels is not only inevitable but in fact the only scientific way to do ethics.

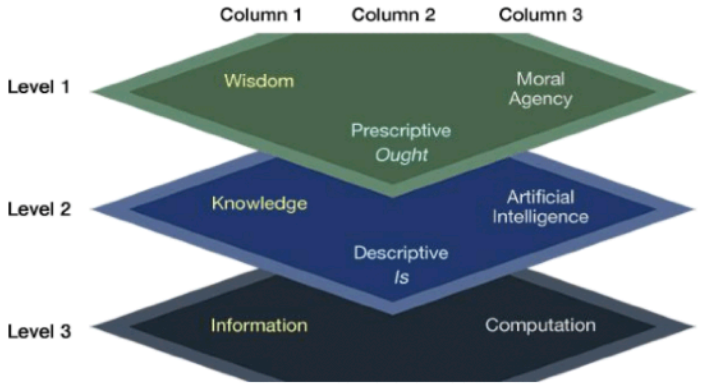

Figure 2: Three Levels: Moral Prescription and Artificial Intelligence

In rebuttal, Hunter and Nedelisky claim, "nearly all of the actual science attempting to deal with morality lands at Level 3 ... since moral disagreements appear not to turn on issues that admit empirical resolution" (2018, 63). Their striking conclusion is that "there are no scientific findings that present claims of either Level One or Level Two status." (2018, 65) Even the new synthesis mentioned above, despite creating lots of excitement in the context of AI ethics, has provided no definitive empirical support for artificially intelligent moral agents or an AI moral science. Even more importantly, for Hunter and Nedelisky, such science cannot tell us what is right and wrong, good or evil, or how we should live.

Bostrom's (2014) worry is that AI is already at Level 2, Knowledge, and accelerating, such that our primary moral question is what we do when it reaches superintelligence at Level 1. Gunkel's worry is how to handle the likely emergence of AI moral agency, Level 3, because morality is computable. Hunter and Nedelisky challenge the new moral science and call us back to a more traditional ethics recognizing the Platonic status of ideas such as the good and virtue. My claim is that none of these programs pose our most pressing moral question at this time.

Science does have a role to play. But it is not clear that a science of morality or an artificially intelligent agent can stand in for the considered, lived understanding of ethics that can only come from an embodied experience of history, literature, poetry, philosophy, and the world's great cultural traditions. We have neither sufficient evidence nor convincing argument, in fact, that ethics is computable.

3. Ethical Robots: Embodied Moral Science

For those who agree that ethics and embodiment are inseparable, and who also object to the idea that a moral science is derivable from scientific theorizing applied to evidence such as neurons, anthropology, and social structure, Wallach and Allen (2009) provide a possible rebuttal of criticisms of the new moral science. Perhaps the pragmatics of robot development, they argue, can free us from unproductive philosophical worries such as free will and the problem of other minds.

Wallach and Allen raise two different worries that seem to impede progress on AI ethics (2009, 58) that can be solved by considering the behavior of robots:

1. Could a robot ever be a moral agent (ontological question)?
2. How could we know that it is such an agent (epistemological question)?

Granting that both questions seem to pose formidable impediments to progress, they suggest we reframe the questions behaviorally, thus sidestepping these unresolvable philosophical questions. Paralleling Turing's (1950) famed sidestepping of philosophical questions, they suggest we simply observe closely how a robot behaves. If it behaves as we would expect an educated, morally aware human to act in morally tricky situations, and does so over time, then we have all the evidence we need to conclude that it is an artificial moral agent, a moral robot.

Given the fact that philosophy often seems to nurture rather than solve such problems, this has appeal. But suppose an exceptional actor agrees to help perpetrate a fraud over a long period of time, in return for a rich

payday. Suppose further for this long period of time this actor's behavior is morally impeccable. At each decision point the actor chooses ethically, cultivating the deep trust the fraud requires. On the basis of this behavior alone, would we be justified in concluding that the actor is moral? On Wallach and Allen's view, "Yes." But when we consider intentions and character, the intuitive answer is "No." This example parallels Searle's (1980) famed "Chinese Room Experiment" which Wallach and Allen call "ground zero in the continuing debate over machine understanding" (2009, 63). If such questions are "ground zero," and we clearly have a problem of induction with the devious actor, sidestepping such philosophical questions—dismissing intuition—invites more Facebook quagmires.

4. The Deceptive Origin of Ethics: The Machiavellian Intelligence Hypothesis

According to the Machiavellian Intelligence Hypothesis (Rowlands 2008, Whiten and Byrne 1997), it is the social world--and the advantages conferred by deception in the social world--that have principally driven increases in primate brain size and intelligence. The hypothesis (MIH) takes sundry forms, but all share the proposition that primate intelligence was driven most by the adaptive complexities of social group dynamics rather than by less socially significant problems such as finding food.

The conventional view is that intelligence led primates to become social since social grouping conferred advantages. According to the MIH, it is the other way about: the competitive dynamics of social grouping accelerated intelligence. That is, primates "became more intelligent because they were social animals" (Rowlands 2008, 60).
But it is not just a matter of being social since most social animals do not become intelligent in the way that primates are intelligent. Canines such as wolves, for example, are intelligent but do not possess primate intelligence. Living in social groups is not just a matter of cooperation but competition.

Why the term "Machiavellian"? Primate societies are distinguished by their interpersonal complexity, which includes the formation of fluid and shifting alliances and coalitions. Within this context, primate social relationships are often manipulative and intentionally *deceptive* at sophisticated levels (Whiten and Byrne 1988). Manipulating, exploiting, but most of all deceiving peers generates additional benefits with acceptable costs—if it is done intelligently. Consequently, the label "Machiavellian intelligence" passed into common usage across a variety of related disciplines.

The origin of ethics, on this sociobiological approach, originates in the competitive advantages offered by deceptive behavior such as lying. Human natural history, for the MIH, was deeply conditioned by the ability of individual primates to deceive better than other deceivers; human culture, art and even science was therefore shaped by our proclivity for deception. If this seems remote from AI technology, Rini (2019) argues "deepfake technology," manufactured video of events that never happened, reduces the trustworthiness of video from a rough equivalence to first-hand observation to dubious "testimony." Primate lying will vitiate sophisticated technology.

If human ethics originated largely in the primate proclivity for deception, a question arises for functionalism in general and moral science in particular. Specifically, can moral science functionally duplicate the path taken by primates from skilled social deception to ethical awareness? The MIH argues that it is the complex mix of learning, socialization and cultural evolution that provides necessary conditions for ethics to arise. We have no evidence that a simulation of this unprecedented complexity is computationally feasible.

4.1 Denying General Intelligence

What we are left with is not a general intelligence that theorizes abstractly about meta-ethics but rather a biological account that claims that ethics originates in the specific primate behavior, with all the contingencies that represents. We can reject classic ethical theories, including Platonism, in the quest for a moral science—and still see that ethics was baked in at the evolutionary beginning. How do we functionally replicate such complexity?

5. Glymour's Rebuttal: the Neuron Replacement Thought Experiment

Russell and Norvig (2015) provide an extended discussion of Glymour's "Brain Replacement Thought Experiment," which defends a functionalist understanding of intelligence and consciousness and may offer the need-

ed rebuttal of the last section. The functionalist understands a mental state to be any intermediate causal structure which produces the right outputs from given inputs.

In the thought experiment, illustrated in Figure 3, the individual neurons in a human brain are replaced, one at a time, until all biological neurons are replaced with artificial neurons. The question is whether the replacement slowly dissolves the person's consciousness.

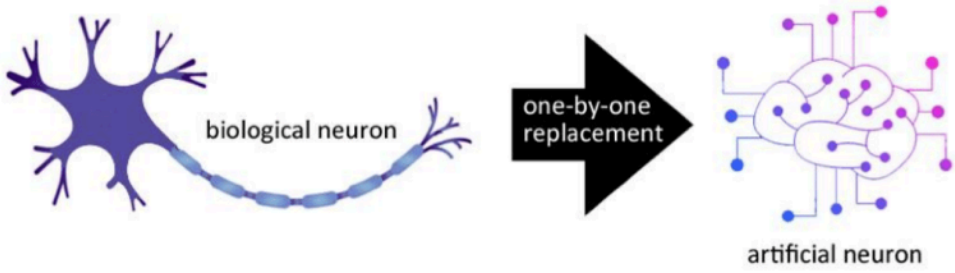

Figure 3: One-By-One Replacement of Biological with Artificial Neurons

Let's assume the best, that the external behavior remains unchanged and we have a good conversation, with consciousness evidently still intact. Russell and Norvig point out that the noted roboticist Hans Moravec accepts the thought experiment as a confirmation of functionalism (brains are not essential to either intelligence or consciousness) but the distinguished philosopher Searle will argue that consciousness vanishes. Russell and Norvig observe, ironically, that "we have a direct clash of intuitions" (2015, 1044).

5.1 Quest for Non-Intuitive Functionalism

Russell and Norvig's argument is that if the person reports an unchanged conscious experience we must agree that we have a confirmation of functionalism. If we are to make more progress, we must agree nothing has been lost since "we have replicated the functional properties of a normal human brain … we must have an explanation of … consciousness … [and] this explanation must also apply to the real brain, which has the same functional properties" (2015, 1045). Importantly, we no longer need appeal to non-physical, non-empirical terms such as "mind," "consciousness," and "intuition." They enthusiastically report that converging research by philosophical physicalists, cognitive science functionalists, and AI proponents will eradicate such "soft" concepts.

But Russell and Norvig move too quickly. As we saw above, there is a compelling case to be made that human intelligence is the product of the complex interplay between biological and social evolution. It is uncontroversial that the human brain is the most complex biological phenomenon we know (Bassett and Gazzaniga 2011). There is insufficient evidence and argument, therefore, to sustain the conclusion that we are close to functionally simulating intuition, mind and consciousness. Dismissing such challenges with a promise of future research is unconvincing.

6. The Replicability Crisis: The New Immoral Science

Science is facing a "replicability crisis" (sometimes "reproducibility crisis") where, it is claimed, more than two-thirds of researchers have tried and failed to reproduce published results. Tim Errington directs The Reproducibility Project, which attempted to repeat the results reported in five pivotal cancer studies. "The idea here is to take a bunch of experiments and to try and do the exact same thing to see if we can get the same results" (Feilden 2017). After research carefully attempting replication over several years (the project was launched in 2011), the team was able to confirm just two of the five original results.

If testing hypotheses by experiment is the heart of scientific method, the replicability of experimental test results is the heart of the integrity of the practice of science. Presumably, authors of refereed journal articles incur the ethical obligation to double check their methodologies and results, and make it clear how others can get the same results. The "replicability crisis" names the growing alarm in the scientific community that this is not the case.

Ioannidis (2005) makes the startling claim that "There is increasing concern that most current published research findings are false." Ioannidis argues that, buttressed with evidence that modest levels of researcher bias, imperfect research techniques, and the tendency to focus on new ideas rather than conventional theories, researchers will generate wrong results most of the time. Expressed simply, if a researcher is attracted to unconventional ideas and is motivated to argue they are correct, and given discretion in the interpretation of the experimental data, the researcher will attempt to conclude the unconventional theories are confirmed by the experimental evidence. Remember that all researchers are primates, with a biologically honed propensity to deceive.

In fact, scientific theories cannot be definitively confirmed by any amount of experimental data. Philosophers of science have understood this for a long time, naming it the problem of "underdetermination" (Stanford 2017). Contrary to popular supposition, then, data alone can never confirm a theory true or false. The reproducibility crisis in natural science means its ethics problem is at least as severe as that of technology and artificial intelligence. As a result, the quest for a "moral science" is ironically morally compromised by science itself.

7. Coarse Graining: When Simulation Lends Itself to Deception

Last, we have a related problem even with the situated moral robots endorsed by Wallach and Allen (2009). Situating robots and enabling them to interact in real time, they point out, need not require the impossibly complex world model of traditional AI. Situated, continuously interacting robots mitigate this classical AI problem, but robots still must interact with an environment that remains a complex adaptive system. Consider Figure 4.

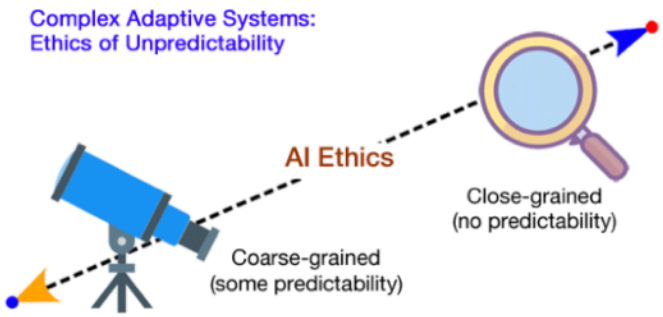

Figure 4: AI and the Ethics of Scientific Unpredictability

The simulation of complex adaptive systems is difficult because such systems, close up, are characterized by computational irreducibility (Wolfram 2002) and are therefore inherently unpredictable. As illustrated in Figure 4, we can gain some coarse-grained predictability by ignoring some local detail of a complex adaptive system. The temptation to trade accuracy for predictability is great and the difference between trustworthy simulation and intentional deception in complex systems is difficult to detect: the line between simulation and deception is thin in multiple ways. This may explain part of the Replicability Crisis since new insights into complex adaptive systems are difficult without knowingly coarse-graining the behavior of the system. It is typically human to find the evidence we are looking for, encouraging the quintessential primate temptation to deceive.

8. Conclusion and Outlook

This paper has argued we have insufficient grounds to conclude that the principal AI ethics problem is that an advancing AI superintelligence will harm humanity, as Bostrom warns, or that soon-to-be-built AI programs and robots will become enslaved moral agents, as Gunkel, Wallach and Allen predict. Instead, the principal AI ethics problem is already here and more practically addressable. Expecting people trained in AI to recognize and handle well the many subtle dimensions of AI ethics without immersion in ethics is equivalent to expecting Hilbert-style formal systems to be able to prove all mathematically true claims. Of course, Gödel proved this impossible. Pham (2019) writes, "when people learn to code they should learn about ethics [and] humanities … then perhaps they'll be more prepared to predict the unintended consequences of their work." We need a Gödel of ethics to show why AI training with only superficial ethical education is the clear-and-present ethics danger facing humanity today.

Specifically, computer science and artificial intelligence curricula should include sustained, close study of ethics, both contemporary ethics and ethical works from antiquity, as well as from a variety of cultures. Ethical judgment resists capture in theory and we have little evidence it is computable. Instead, it is an embodied human sensibility, a set of dispositions and interpersonal skills, nurtured in a mix of socialization, ethics study and lived ethical dilemmas.

The sobering news is that this significant change in computer science curricula is urgent. The encouraging news is that we know how to make this change and how to do it now.

References

Bassett, D. and Gazzaniga, M. (2011) "Understanding Complexity in the Human Brain," *Trends Cog. Sci.* 2011 May, 200-209.

Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies,* Oxford University Press, Oxford UK.

Bringsjord, S.and Govindarajulu, N., "Artificial Intelligence", *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition), Edward N. Zalta (ed.).

Crockett, L. (2002) "Fundamental Issues in Honors Teaching," *Teaching and Learning in Honors*, National Collegiate Honors Council , ed. Fuiks and Clark, 2002, 21-32. http://digitalcommons.unl.edu/nchcmono/9/.

Crockett, L. (1994) *The Turing Test and the Frame Problem*, Ablex Publishing, Norwood N. J.

Descartes, R. (1637), in Haldane, E. and Ross, G., translators, 1911, *The Philosophical Works of Descartes*, Volume 1, Cambridge, UK: Cambridge University Press.

Dreyfus, H. (1972) *What Computers Can't Do*, Cambridge, MA: MIT Press.

Dreyfus, H. (1992) *What Computers Still Can't Do*, Cambridge, MA: MIT Press.

Feilden, T. (2017) "Most Scientists 'Can't Replicate Studies By Their Peers'" Feb. 22, 2017. https://www.bbc.com/news/science-environment-39054778.

Gödel, K. (1931) "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I", *Monatshefte für Mathematik und Physik*, 38: 173–198.

Grimes, W. (2017) "Hubert L. Dreyfus, Philosopher of the Limits of Computers," *New York Times,* May 2, 2017.

Gunkel, D. (2012) *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*, MIT Press, Cambridge, Mass.

Hamilton, I. (2019) "Shame On Mark Zuckerberg." *https://business*insider.com/mark-zuckerberg-slammed-no-evidence-to-lawmakers-in-canada-2019-5.

Hunter, J. and Nedelisky, P. (2018) *Science and the Good,* Yale University Press, New Haven, CT.

Ioannidis, J. (2005) "Why Most Published Research Findings Are False" PLOS Med. Aug. 30 2005. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182327/.

McFarland, M. (2014) Elon Musk: "With artifical intelligence we are summoning the demon." *The Washington Post,* Oct. 24, 2014.

Moore, G. (1903) *Principia Ethica,* Cambridge University Press, Cambridge, UK.

Osborne, H. (2017) "Stephen Hawking AI Warning." *Newsweek,* Nov. 7, 2017.

Pham, K. (2019) "Want to Fix Big Tech? Change What Classes Are Required for a Computer Science Degree", *Fast Company*, 5.28.19. https://www.fastcompany.com/90355969/want-to-fix-big-tech-change-what-classes-are-required-for-a-computer-science-degree.

Pinker, S. (2018) *Enlightenment Now*, Viking, New York, NY.

Rini, R. (2019) "Deepfakes are coming. We can no longer believe what we see," *New York Times*, June 10, 2019.

Rowlands, M. (2008) *The Philosopher and the Wolf*, Pegasus Books, New York, NY.

Russell, S. and Norvig, P. (2015) *Artificial Intelligence: A Modern Approach*, Pearson Education, Tamil, Nadu, India.

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

Stanford, Kyle (2017) "Underdetermination of Scientific Theory", *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), https://plato.stanford.edu/archives/win2017/entries/scientific-underdetermination/.

Turing, A. (1950) "Computing Machinery and Intelligence," *Mind*, LIX: 433–460.

Turing, A. (1936) "On Computable Numbers with Applications to the Entscheidung-Problem," *Proceedings of the London Mathematical Society,* 42: 230–265.

Wallach, W. and Allen, C. (2009) *Moral Machines*, Oxford University Press, New York, NY.

Whiten, A. and Byrne, R., ed. (1988) *Machiavellian Intelligence,* Clarendon Press, UK.

Whiten, A. and Byrne, R., ed. (1997) *Machiavellian Intelligence II*, Cambridge University Press, Cambridge, U.K.

Wolfram, S. (2002) "Computational Irreducibility" Wolfram Mathworld http://mathworld.wolfram.com/PrincipleofComputationalIrreducibility.html.