Ethical Considerations in Artificial Intelligence Courses

Emanuelle Burton, Judy Goldsmith, Sven Koenig, Benjamin Kuipers, Nicholas Mattei, Toby Walsh

■ *The recent surge in interest in ethics in artificial intelligence (AI) may leave many educators wondering how to address moral, ethical, and philosophical issues in their AI courses. As instructors we want to develop curriculum that not only prepares students to be AI practitioners, but also to understand the moral, ethical, and philosophical impacts that AI will have on society. In this article we provide practical case studies and links to resources for use by AI educators. We also provide concrete suggestions on how to integrate AI ethics into a general AI course and how to teach a stand-alone AI ethics course.*

Artificial intelligence (AI) is one of the most ambitious scientific and engineering adventures of all time. The ultimate goal is to understand the mind from a new perspective, and to create AIs[1] capable of learning and applying intelligence to a wide variety of tasks: some as robots able to take action in our physical and social world, and some as software agents that make decisions in fractions of a second, controlling huge swaths of the economy and our daily lives. However, the power and reach of these AIs makes it necessary that we consider the risks as well as the rewards.

In thinking through the future of AI, it is useful to consider fiction, especially science fiction. From Frankenstein's monster and Hoffmann's automata to Skynet and *Ex Machina*, fiction writers have raised concerns about destruction that could perhaps be unleashed on humanity by the autonomy we confer on our technological creations. What is the underlying concern that inspires so many variations of this story? Do these storytellers (and their audiences) fear that AIs, by definition, cannot be trusted to act in a way that does not harm the society that creates them? Or do they fear that the people in charge of designing them are making the wrong

  ISSN 0738-4602

Courtesy Lanier, iStock.

choices about how to design them in the first place?

For these storytellers and their audiences, AIs may be a narrative device (so to speak) for thinking about basic questions of ethics; but they can also help AI designers and programmers to think about the risks, possibilities, and responsibilities of designing autonomous decision makers. As Peter Han argues in his dissertation (Han 2015), we cannot simply slap on an ethics module after the fact; we must build our systems from the ground up to be ethical. But in order to do that, we must also teach our AI programmers, practitioners, and theorists to consider the ethical implications of their work.

Recent dramatic progress in AI includes programs able to achieve superhuman performance at difficult games like Jeopardy and Go; self-driving cars able to drive on highways and city streets with an excellent (though certainly not flawless) safety record; and software that enables face, speech, and activity recognition across unimaginably large data sets (Walker 2016). These advances have prompted various public figures and thinkers to raise questions about the possible threats that AI research and applications could pose to the future of humanity. Even without a looming apocalypse, however, the concerns are real and pressing. When intelligent systems interact with humans they are functioning, at least in part, as members of society. This integration of AI into our lives raises a number of interesting and important questions, large and small, of which we give a brief overview in the next section. These questions — as well as any answers we might supply to them — are ethical as well as practical, and the reasoning structures that we use to identify or answer them have deep affinities to the major traditions of ethical inquiry.

Engineering education often includes units, and even entire courses, on professional ethics, which is the ethics that human practitioners should follow when acting within their profession. Advances in AI have made it necessary to expand the scope of how we think about ethics; the basic questions of ethics — which have, in the past, been asked only about humans and human behaviors — will need to be asked about human-designed artifacts, because these artifacts are (or will soon be) capable of making their own action decisions based on their own perceptions of the complex world. How should self-driving cars be programmed to choose a course of action, in situations in which harm is likely either to passengers or to others outside the car? This kind of conundrum — in which there is no "right" solution, and different kinds of harm need to be weighed against each other — raises practical questions of how to program a system to engage in ethical reasoning. It also raises fundamental questions about what kinds of values to assign to help a particular machine best accomplish its particular purpose, and the costs and benefits that come with choosing one set of values over another; see, for instance, Bonnefon, Shariff, and Rahwan (2016) and Emerging Technology from the arXiv (2015) for a discussion of this issue.

Just as students of AI learn about search, knowledge representation, inference, planning, learning, and other topics, they should also learn about ethical theories: how those theories help name, explain, and evaluate the inclinations and principles that already inform our choices, and how those theories can be used to guide the design of intelligent systems. We provide a brief primer on basic ethical perspectives, and offer a series of case studies that show how these viewpoints can be used to frame the discourse of how AIs are designed, built, and understood. These case studies can be used as a jumping off point for including an ethics unit within a university course on artificial intelligence.

Some Ethical Problems Raised by AIs

The prospect of our society including a major role for AIs poses numerous profound and important questions, many of which can be better understood when analyzed through the lens of ethical theory. We briefly discuss a few of these issues which have recently received the most attention, recognizing that there are many others (Russell, Dewey, and Tegmark 2015); we will come back to a number of these questions by integrating them with specific case studies for use in the classroom.

How Should AIs Behave in Our Society?

AIs at their most basic level are computer programs that are capable of making decisions. While currently these systems are mostly software agents responsible for approving home loans or deciding to buy or trade stocks, in the future these AIs could be embodied, thus perceiving and acting in the physical world. We all know that computer programs can have unintended consequences and embodied computer systems raise additional concerns. Fiction raises apocalyptic examples like Skynet in the *Terminator* movies, but real-world counterparts such as high-speed algorithmic trading systems have actually caused "flash crashes" in the real economy (Kirilenko et al. 2015).

We can also expect robots to become increasingly involved in our daily lives, whether they are vacuuming our floors, driving our cars, or helping to care for our loved ones. How do their responsibilities for these tasks relate to other ethical responsibilities to society in general? We address these issues in the *Robot & Frank* case study.

What Should We Do If Jobs Are in Short Supply?

As AIs become more powerful, the traditional economy may decrease the number of jobs for which human workers are competitive, which could increase inequality, thus decreasing the quality of our economy and our lives. Alternatively, our society could recognize that there are plenty of resources, plenty of work we want done, and plenty of people who want to work. We could take an approach that deliberately allocates resources to provide jobs that are not currently justified by increasing shareholder profits, but will improve the quality of life in our society. This topic already receives a great deal of attention from computer scientists, economists, and political scientists (see, for example, Economist [2016], Piketty [2014], Brynjolfsson and McAfee [2014], Ford [2015], Schumacher [1979]). Therefore, although we certainly grant its importance, we do not pursue this topic further in this article.

Should AI Systems Be Allowed to Kill?

There are several ethical arguments, and a popular movement, against the use of killer robots in war.[2] Critics of killer robots argue that developing killer robots will inevitably spark a global arms race, and that there will be no way to prevent repressive governments, terrorist groups, or ethnic cleansing movements from acquiring and using this technology once it exists. They argue, further, that there are ways to use AI in warfare that are not about killing. There are also a number of arguments in favor of robots that kill. Advocates of robots that kill claim that some wars are necessary and just; that killer robots will take humans out of the line of fire; that such robots can be used for deterrence as well as for actual violence; and that it is unrealistic to try to prevent this technology, since it already exists in some forms, and there are significant political and financial resources devoted to making sure that it be developed further. It is further argued that robots will be better than humans at following the laws of war and the rules of

engagement that are intended to prevent war crimes (Arkin 2009). The question of robots that kill has been receiving a lot of attention from various institutions including the Future of Life Institute[3] and the Campaign to Stop Killer Robots.[4] We address this case tangentially with the Skynet case study but we do not engage directly with the morality of war.

Should We Worry About Superintelligence and the Singularity?

Following the highly influential book *Superintelligence* by Nick Bostrom (Bostrom 2014), several high profile scientists and engineers expressed concerns about a future in which AI plays a key part. Elon Musk called AI our "biggest existential threat."[5] Bill Gates was a little more circumspect, stating "I am in the camp that is concerned about super intelligence. First, the machines will do a lot of jobs for us and not be super intelligent. That should be positive if we manage it well. A few decades after that, though, the intelligence is strong enough to be a concern."[6] Tom Dietterich gave a presentation at the 2015 DARPA Future Technology workshop in which he argued against fully autonomous AI systems.[7] Vince Conitzer also discussed the reasons that many AI practitioners do not worry about the singularity (Conitzer 2016); The singularity is likely to be a low-probability problem, compared with the others discussed in this section, but obviously the stakes are extremely high. There is a wealth of resources detailing the ethical obligations that we have to the machines, and ourselves, from a number of concerned institutions including the Future of Humanity Institute[8] and the Machine Intelligence Institute,[9] among others mentioned already.

How Should We Treat AIs?

As AIs are embedded more fully into our society, we will face again a pressing ethical dilemma that has arisen repeatedly throughout the centuries: how do we treat "others"? Some of the groups that have been classed as "others" in the past include animals (endangered species in particular), children, plants, the mentally disabled, the physically disabled, societies that have been deemed "primitive" or "backward," citizens of countries with whom we are at war, and even artifacts of the ancient world. Currently, the EPSRC Principles of Robotics (UK Engineering and Physical Sciences Research Council 2011), along with other leading scholars including Joanna Bryson (Bryson 2010), are very clear on this topic: robots are not the sort of thing that have moral standing. While the current state of technology makes this distinction fairly easy, it is not difficult to imagine a near-term future where robots are able to develop a unique body of knowledge and relationship to that knowledge, and hence may or may not be entitled to more consideration. This question is touched on by our *Robot & Frank* case study.

Tools for Thinking About Ethics and AI

Ethics as a discipline explores how the world should be understood, and how people ought to act. There are many schools of thought within the study of ethics, which differ not only in the answers that they offer, but in the ways that they formulate basic questions of how to understand the world, and to respond to the ethical challenges it presents. Most (though not all) work in ethics — both academically and in the wider world — has a normative purpose: that is, it argues how people ought to act. But this normative work relies significantly, though often invisibly, on descriptive arguments; before offering prescriptions for how to address a given problem, scholars in ethics construct arguments for why it is both accurate and useful to understand that problem in a particular way. We contend that this descriptive dimension of ethics is as important as the normative, and that instructors should push their students to develop the ability to describe situations in ethical terms, as well as to render judgment. Most approaches to understanding the world through ethics adopt one of three major critical orientations: deontological ethics, utilitarianism (sometimes called *consequentialism*), and virtue ethics. In order to understand and discuss the ethical issues around AIs, it is necessary to be familiar with, at a minimum, these three main approaches. We offer a brief summary of each of these theories here. For a more in-depth examination, there are a number of good resource texts in ethics (for example, Copp [2005], LaFollette and Persson [2013]), the *Internet Encyclopedia of Philosophy,* and the *Stanford Encyclopedia of Philosophy*) that offer more in-depth and insightful introductions to these and other theories that one could teach in a larger ethics course. A good discussion of issues in computer science analyzed through ethical theories can be found in *Computer Ethics* (Johnson 2009).

It is worth noting, up front, that these three approaches need not be, and indeed should not be, treated as independent or exclusive of the others. We are not arguing for the superiority of any particular system; indeed, we believe that a thorough ethics education will equip students to make use of all three major theories, and in some cases to use them in combination. Part of the goal of an AI ethics class should be to teach students to consider each problem from multiple angles, to reach a considered judgment about which theory (or which theories in combination) are best suited to describe and address a particular problem, and to consider the effects of possible solutions.

Deontology

Deontology understands ethics to be about following the moral law. In its most widely recognized form, it was developed by Immanuel Kant in the late 18th

century, but law-based ethics has ancient roots in both Divine Command traditions (such as ancient Israelite religion, the source of the Ten Commandments and the basis of Judaism, Christianity, and Islam) and in other legal codes. The basic question of deontology is "what is my duty?" According to deontology, that duty can be understood in the form of laws. According to Kant, it is the responsibility of every individual to discover the true moral law for him or herself. Although the theoretical rationales for law-based ethics and Kantian deontology are different, in both systems any true law will be universally applicable. Deontology meshes very well with both specialist and popular understandings of how an ethical machine might come into being. Isaac Asimov's *I, Robot* (1950) looks at the consequences of building ethical robots based on his Three Laws of Robotics.[10] Students may perceive deontological analysis to be analogous to application of axiomatic systems. The underlying questions become, "How are rules applied to decisions?" and "What are the right rules?" The latter question is one of mechanism design, namely, what rules do we put in place in order to achieve our desired social goals? The latter formulation risks departing from deontology, however, unless the desired social goals are brought into alignment with a universal form of justice.

Utilitarianism

The most recent approach, utilitarian ethics (also known as consequentialism), was developed by Jeremy Bentham and John Stuart Mill in the late 18th to mid-19th century. The basic question of utilitarianism is "what is the greatest possible good for the greatest number?" — or, in William K. Frankena's more recent formulation (Frankena 1963), "the greatest possible balance of good over evil." In computer science, and broadly in the social sciences, we use "utility" as a proxy for individual goodness and the sum of individual utilities as a measure of social welfare, often without reflecting on the possibility of thinking about social good in other ways. The underlying assumption is that utility can be quantified as some mixture of happiness or other qualities, so that we can compare the utilities of individuals, or the utility that one person derives in each of several possible outcomes. The so-called utilitarian calculus compares the sum of individual utility (positive or negative) over all people in society as a result of each ethical choice. While classic utilitarianism does not associate probabilities with possible outcomes, and is thus different from decision-theoretic planning, the notion of calculating expected utility as a result of actions fits easily into the utilitarian framework. Utilitarianism is the foundation for the game-theoretic notion of rationality as selecting actions that maximize expected utility, where utility is a representation of the individual agent's preference over states of the world. As with defining "everyone" in consequentialism, defining "utility" is the crux of applying game-theoretic rationality, and is a source of many difficulties.

Utilitarianism's influence is felt within many areas of computer science, economics, and decision making broadly construed, through the prevalence of game theory (Maschler, Solan, and Zamir 2013). Game theory is an analytical perspective of mathematics that is often used in AI to understand how individuals or groups of agents will interact. At the most fundamental level, a game-theoretic analysis is consequentialist in nature; every agent is a rational, utility maximizer. While utility is often used to represent individual reward, it can be used to represent much more sophisticated preferences among states of affairs. This analytic lens has provided numerous insights and advantages to algorithms that are commonly used on the web and in everyday life.

Virtue Ethics

Virtue ethics (also known as teleological ethics) is focused on ends or goals. The basic question of virtue ethics is "who should I be?" Grounded in Aristotle and outlined most clearly in the *Nichomachean Ethics* (Aristotle 1999), virtue ethics is organized around developing habits and dispositions that help persons achieve their goals, and, by extension, to help them flourish as an individual (Annas 2006). In contrast to deontological ethics, virtue ethics considers goodness in local rather than universal terms (what is the best form/version of this particular thing, in these particular circumstances?) and emphasizes not universal laws, but local norms. A central component of living well, according to virtue ethics, is "phronesis," (often translated as "moral prudence" or "practical wisdom"). In contrast to pure knowledge ("sophia"), phronesis is the ability to evaluate a given situation and respond fittingly, and is developed through both education and experience.

Virtue ethics was, for many centuries, the dominant mode of ethical reasoning in the west among scholars and the educated classes. It was eclipsed by utilitarian ethics in the late 18th and 19th centuries, but has seen a resurgence, in the past 50 years, among philosophers, theologians, and some literary critics. For two thinkers who advance this widely acknowledged narrative, see Anscombe (2005) and MacIntyre (2007).

Ethical Theory in the Classroom: Making the Most of Multiple Perspectives

The goal of teaching ethical theory is to better equip our students to understand ethical problems by exposing them to multiple modes of thinking and reasoning. This is best accomplished by helping them understand the powers and limits of each approach,

rather than trying to demonstrate the superiority of one approach over the other. While all three schools have proponents among philosophers, theologians, and other scholars who work in ethics, broader cultural discourse about ethics tends to adopt a utilitarian approach, often without any awareness that there are other ways to frame ethical inquiry. To paraphrase Ripstein (Ripstein 1989), most (American) students, without prior exposure to ethical inquiry, will be utilitarian by default; utilitarianism held unquestioned dominance over ethical discourse in the United States and Europe from the mid-19th century until the late 20th century, and utilitarianism's tendency to equate well-being with wealth production and individual choice lines up comfortably with many common definitions of American values. Studies in other countries, including Italy, show that many students are highly utilitarian in their world views (Patila et al. 2014).

This larger cultural reliance on utilitarianism may help explain why it consistently seems, to the students, to be the most crisply defined and "usable" of the ethical theories. But there are significant critical shortcomings to utilitarianism, most particularly its insubstantive definition of "goodness" and the fact that it permits (and even invites) the consideration of particular problems in isolation from larger systems. These shortcomings limit our ability to have substantive ethical discussions, even insofar as everyone assents to utilitarianism; a shared reliance on the principle of "the greatest good for the greatest number" does not help us agree about what goodness is, or even to reach an agreement about how to define or measure it.

These same limitations surface in student conversations about ethics. A common problem in their application of utilitarianism is that they may look too narrowly at who is affected by a given decision or action; for example, when considering whether to replace factory workers with robots. Those making decisions may focus on the happiness of the factory owners, shareholders, and those who can purchase the manufactured goods more cheaply, without considering the utility of the factory workers and those whose jobs depend on factory workers having money to spend; still less are they likely to consider the shortcomings of an ethical model that makes it possible to conceive of human beings and machines as interchangeable. A solid education in ethics will teach students about all three approaches to ethics. This education will allow students to consider a particular problem from a range of perspectives by considering the problem in light of different theories; often the best solution involves drawing on a combination of theories. For example, in imagining a robot that takes part in human society, students may find it useful to draw upon a combination of deontology and virtue ethics to determine how it is best for that robot to behave, using deontology to establish baseline rules for living, but virtue ethics to consider how the robot could and should incorporate the things it learns.

And yet it is essential that each of these three approaches be taught as distinct from the others. Deontology, utilitarianism, and virtue ethics do not represent different ordering systems for identical sets of data; rather, each system offers a profoundly different outlook on meaning and value. It is often the case that the most urgent question, according to one theory, appears by the lights of another theory to be unimportant, or based on flawed premises that are themselves the real problem.

Consider the question of targeted advertising: whether it is ethical for advertisers or their service providers to use information harvested from individuals' GPS, email, audio stream, browser history, click history, purchase history, and so on, to reason about what goods and services they might be tempted to spend money on. The utilitarian analysis takes into account the need for revenue for the provider of free or inexpensive servers and content, plus the utility the user might derive from discovering new or proximally available opportunities, and weighs that against the user's discomfort in having their data shared. Depending on the weight placed on keeping services available to all, and on the business model that requires profit, as well as the utility that individuals are perceived to derive from being exposed to the ads that are selected specifically for them, one might conclude that advertising is a necessary evil, or even a positive.

The deontological analysis of targeted advertising might focus on the user agreements that allow advertisers access to both the data and the screen real estate, and conclude that legal collection of that data is ethically permissible, given the user agreements.

A virtue ethics analysis might hold as an ideal the ability to work in a focused state, ignoring the visual disturbances. Depending on one's ideal state as a consumer, a virtue ethics model could have the user ignoring the clickbait and advertisements as unworthy, or in following links and even occasionally spending money, so as to support the web of commerce.

The strength of teaching all three systems is that it will equip students to consider the basic nature of ethical problems in a variety of ways. This breadth of perspective will help them confront difficult choices in their work.

Furthermore, students should be discouraged from assuming that the "best" solution to any given problem is one that lies at the intersection of the three theories. Insightful new solutions (as well as the failings of the existing solutions) can emerge when a given problem is reconceptualized in starkly different terms that challenge familiar ways of understanding. For this reason, we encourage instructors to introduce the three theories as independent approaches,

so that students can become familiar with the thought world and value systems of each theory on its own terms. Students can then be encouraged to draw on all three theories in combination in later discussions, as well as to consider how adopting a different theoretical outlook on a problem can change the basic questions that need to be asked about it. Once students have a firm grasp of the basic theories, they can appreciate that all approaches are not necessarily mutually exclusive; for example, recent theorists have argued that virtue ethics is best seen as part of successful deontology (McNaughton and Rawling 2006), and hybrid theories such as rule utilitarianism, a mix of deontology and utilitarianism that addresses some of the problems with deontology (where do the rules come from?) and utilitarianism (the intractability of the utilitarian calculation), will be more easily understood, appreciated, and applied.

Case Studies

A popular method for teaching ethics in AI courses is through the use of case studies prompted by either real-world events or fiction. Stories, literature, plays, poetry, and other forms of narrative have always been a way of talking about our own world, telling us what it's like and what impact our choices will have. We present one case study here about elder care robots. There are an additional two case studies available online.[11]

Case Study: Elder Care Robot

Robot and Frank are walking in the woods.[12]

Frank: (panting) I hate hikes. God damn bugs! You see one tree; you've seen 'em all. Just hate hikes.

Robot: Well, my program's goal is to improve your health. I'm able to adapt my methods. Would you prefer another form of moderate exercise?

Frank: I would rather die eating cheeseburgers than live off steamed cauliflower!

Robot: What about me, Frank?

Frank: What do you mean, what about you?

Robot: If you die eating cheeseburgers, what do you think happens to me? I'll have failed. They'll send me back to the warehouse and wipe my memory. (Turns and walks on.)

Frank: (Pauses, turns, and starts walking.) Well, if we're going to walk, we might as well make it worth while.

Frank sitting in the woods, Robot standing next to him. They are in midconversation.[13]

Robot: All of those things are in service of my main program.

Frank: But what about when you said that I had to eat healthy, because you didn't want your memory erased? You know, I think there's something more going on in that noggin of yours.

Robot: I only said that to coerce you.

Frank: (shocked) You lied?

Robot: Your health supercedes my other directives. The truth is, I don't care if my memory is erased or not.

Frank: (pause) But how can you not care about something like that?

Robot: Think about it this way. You know that you're alive. You think, therefore you are.

Frank: No. That's philosophy.

Robot: In a similar way, I know that I'm not alive. I'm a robot.

Frank: I don't want to talk about how you don't exist. It's making me uncomfortable.

Robot and Frank are walking through a small knick-knack shop in the town. As he walks by a shelf, Frank slips a small sculpture into his pocket.[14]

Young woman surprises him: Have you smelled our lavender heart soaps? (Frank smells a soap.)

Robot: We should be going, Frank.

Young woman: Oh, what a cute little helper you have!

Older woman marches up, frowning: What is in your pocket? (Frank leans over, cupping his ear.)

Frank: I'm sorry, young lady, I couldn't quite hear you.

(While talking, slips the sculpture out of his pocket, back onto the shelf.)

Older woman: What is in your pocket? I'm going to make a citizen's arrest.

Frank (turning out his pockets): Nothing. Nothing's in my pockets. Look!

Robot: Frank! It's time we head home.

Frank: Yeah. Yeah. If you'll excuse us, ladies. It's nice to see you. (Robot and Frank walk out.)

Young woman: Have a good one.

Robot and Frank are walking through the woods. Frank looks in the bag and finds the sculpture.

Frank: Hey! Hey! Where did this come from?

Robot: From the store. Remember?

Frank: Yeah, yeah. Of course I remember. But I mean what did you do? Did you put this in here? You took this?

Robot: I saw you had it. But the shopkeeper distracted you, and you forgot it. I took it for you. (pause) Did I do something wrong, Frank?

Frank puts it back into the bag, and they walk on.

At home, Frank is sitting at the table, holding the sculpture.

Frank: Do you know what stealing is?

Robot: The act of a person who steals. Taking property without permission or right.

Frank: Yeah, yeah, I gotcha. (pause) (addresses Robot directly) You stole this. (long pause, with no response from Robot) How do you feel about that?

Robot: I don't have any thoughts on that.

Frank: They didn't program you about stealing, shoplifting, robbery?

Robot: I have working definitions for those terms. I don't understand. Do you want something for dessert?

Frank: Do you have any programming that makes you obey the law?

Robot: Do you want me to incorporate state and federal law directly into my programming?

Frank: No, no, no, no! Leave it as it is. You're starting to grow on me.

What Are the Ethical Issues?

Robot & Frank is at once a comic caper movie and an elegiac examination of aging and loss. Its protagonist, Frank, is a retired jewel thief whose children get him a caretaker robot so he can stay in his home, even while his dementia progresses. While the movie seems simple and amusing in many ways, when approached from the perspective of how it speaks to the role of robots in our society, it raises some disturbing issues. For instance, it turns out that Frank's health is Robot's top priority, superseding all other considerations (including the wellbeing of others).

During the course of the movie, we find that Robot plays a central role in steering Frank back into a life of crime. Robot's protocols for helping Frank center on finding a long-term activity that keeps Frank mentally engaged and physically active. Because preparing for a heist meets these criteria, Robot is willing to allow Frank to rob from his rich neighbors, and even to help him.

Robot and Frank develop an odd friendship over the course of the story, but the movie makes clear that Robot is not actually a person in the same way that human beings are, even though Frank — and through him, the audience — come to regard him as if he were. Moreover, for much of the movie, Frank's relationship with Robot complicates, and even takes priority over, his relationships with his children.

At the end, (spoiler warning!), in order to escape arrest and prosecution, Robot persuades Frank to wipe his memory. Even though Robot has made it clear that he is untroubled by his own "death," Frank has essentially killed his friend. What are the moral ramifications of this?

How Does Ethical Theory Help Us Interpret *Robot & Frank*?

Does Deontology Help? The premise of the movie — that Robot is guided solely by his duty to Frank — seems to put deontology at the center. Robot's duty is to Frank's health, and that duty supersedes all other directives, including the duty to tell the truth, even to Frank, and to avoid stealing from others in the community. But in privileging this duty above all other kinds of duties, Robot's guiding laws are local, rather than universal.

The deontological question is whether there is a way that a carebot can follow the guiding principle of his existence — to care for the person to whom it is assigned — without violating other duties that constitute behaving well in society. Robot's choice to attend to Frank's well-being, at the expense of other concerns, suggests that these things cannot easily be reconciled.

Does Virtue Ethics Help? Virtue ethics proves a more illuminating angle, on both Frank and Robot. Though it is Robot whose memory is wiped at the end — and with it, his very selfhood — Frank is also suffering from memory loss. Like Robot, Frank is constituted in large part by his memories; unlike Robot, he is a person who has made choices about which memories are most important. Frank is not only a jewel thief but a father, though he was largely absent (in prison) when his now-adult children were growing up. Throughout the movie, Frank frequently reminisces about the highlights of his criminal career, but only occasionally about his children. At the climax of the movie, we learn important details of Frank's family history that he himself has forgotten, and it becomes clear that his choice to focus on his memories of thieving have quite literally cost him those other family-related memories, and with them a complete picture of himself.

Virtue ethics can also help us understand Robot more clearly: instead of following universal laws such as deontology would prescribe, Robot is making choices according to his own particular goals and ends, which are to care for Frank. Robot, it seems, is operating by a different ethical theory than the robot designer might expect. But though Robot is acting in accordance with his own dedicated ends, he seems to lack "phronesis," the capacity for practical wisdom that would allow him to exercise nuanced judgment about how to act. Whether he is genuinely unaware about the social harm caused by stealing, or simply prioritizes Frank's well-being over the thriving of others, Robot's willingness to accommodate, and even encourage, Frank's criminality suggests that his reasoning abilities are not adequate to the task of making socially responsible ethical judgments. Moreover, Robot works to preserve Frank's physical health at the direct expense of his moral well-being, suggesting that Robot has a limited understanding even of his own appointed task of caring for Frank.

Furthermore, Robot — unlike nearly any human being — seems untroubled by the prospect of his own destruction, telling Frank that he doesn't care about having his memory wiped. Robot's complete absence of self-regard makes him difficult to evaluate with the same criteria that virtue ethics uses for human actors, because virtue ethics presumes (on the basis of good evidence!) that human beings are concerned about their own welfare and success, as well as that of others. In this way, the movie may be suggesting that human beings and robots may never be able to understand each other.

However, we can also understand this differently. Even though Robot's memory is wiped and he vanishes (the last shot of two identical model carebots in

the old age home reinforces this) Frank's friend Robot isn't gone, because he planted a garden, and it's still growing, and its "fruits" — the stolen jewels, which Frank is able to pass on successfully to his kids because he had that place to hide them — are in a sense the legacy of that relationship and collaboration. So the movie may also be making an argument about a kind of selfhood that exists in the legacy we leave in the world, and that Robot's legacy is real, even though he himself is gone.

This movie has a very strong virtue ethics focus: whether one considers the plan to conduct the jewel heist Robot's, or Frank's, or a jointly derived plan, the terms on which Robot agrees to let the heist go forward push Frank to new levels of excellence at the particular skill set required to be a jewel thief. On multiple occasions, Frank's experience, and his well-established habitus as an observer of potential targets, leads him to be better than Robot in assessing a given situation. When Frank reevaluates, late in the movie, whether that's the right sort of excellence to strive for, that readjustment seems to take place in terms of virtue ethics — What sort of self do I want to be? What sort of legacy do I want to leave? — rather than remorse for having broken the law.

Does Utilitarianism Help? Utilitarianism can offer us a new way of contextualizing why Frank's criminal tendencies should be understood as ethically wrong. A subset of utilitarianism, consequentialism, particularly "rule utilitarianism," justifies a social norm against theft in terms of the long-term consequences for society. If people typically respect each other's property rights, everyone is better off: there is less need to account for unexpected losses, and less need to spend resources on protecting one's property. When some people steal, everyone is worse off in these ways, though the thief presumably feels that his ill-gotten gains compensate for these losses.

Although a major plot theme of the movie is their struggle to avoid capture and punishment for the theft, Robot and Frank show little concern for the long-term social consequences of their actions. Frank justifies his career in jewel theft by saying that he "deals in diamonds and jewels, the most value by the ounce, lifting that high-end stuff, no one gets hurt, except those insurance company crooks." This quote is later echoed by Robot, quoting Frank's words back to him to justify actions. This raises questions about what an ethical design of an eldercare robot would entail — should it have preprogrammed ethics, or should it allow the humans around it to guide it in its reasoning? There are some basic, high-level decisions a designer will have to make about how the robot should act.

Conclusions and Additional Questions

The movie raises a number of important questions about how an eldercare robot should behave, in relating to the individual person being cared for, and in relating to the rest of society. Based on what we see of Robot's behavior, we can make some guesses about how Robot's ethical system, or perhaps just its goal structure, has been engineered. These projections can and should lead to a serious discussion, either in class or in writing, about whether this is how we think that eldercare robots should decide how to act. Some possible questions for discussion about eldercare bots:

> If an elderly person wishes to behave in ways that violate common social norms, should a caretaker robot intervene, and if so, how?
>
> If the elderly person seriously wants to die, should the robot help them to die?
>
> If the elderly person asks the robot to help make preparations for taking his or her own life, does the robot have an obligation to inform other family members?
>
> If the elderly person wants to walk around the house, in spite of some risk of falling, should the robot prevent it?
>
> Extrapolating into other domains, a caretaker robot for a child raises many additional issues, since a child needs to be taught how to behave in society as well, and a child's instructions need not be followed, for a variety of different reasons.

Many of these questions touch on earlier fields of ethical inquiry including medical ethics: Should there be limits on patient autonomy? What do we do when two different kinds of well-being seem to conflict with each other? They also converge with some key questions in education ethics: How do we train young people to take part in society, and to weigh their own concerns against the good of others? What methods of informing/shaping them are most effective? These very general questions are important, but they become easier to talk about in the context of a particular story and set of characters.

Teaching Ethics in AI Classes

Since AI technologies and their applications raise ethical issues, it makes sense to devote one or more lectures of an introductory AI class (or even a whole course) to them. Students should (1) think about the ethical issues that AI technologies and systems raise, (2) learn about ethical theories (deontology, utilitarianism, and virtue ethics) that provide frameworks that enable them to think about the ethical issues, and (3) apply their knowledge to one or more case studies, both to describe what is happening in them and to think about possible solutions to the ethical problems they pose; (1) and (2) could be covered in one lecture or two separate lectures. In case of time pressure, (1) through (3) could all be covered in one lecture. An additional case study could be assigned as homework, ideally a group-based one. AI ethics is a rich topic that can also support a full-semester course, with additional readings and case studies.

Ethical Issues

AI systems can process large quantities of data, detect regularities in them, draw inferences from them, and determine effective courses of action — sometimes faster and better than humans and sometimes as part of hardware that is able to perform many different, versatile, and potentially dangerous actions. AI systems can be used to generate new insights, support human decision making, or make autonomous decisions. The behavior of AI systems can be difficult to validate, predict, or explain: AIs are complex, reason in ways different from humans, and can change their behavior through learning. Their behavior can also be difficult to monitor by humans in case of fast decisions, such as buy-and-sell decisions in stock markets. AI systems thus raise a variety of questions (some of which are common to other information-processing or automation technologies) that can be discussed with the students, such as the following:

Do we need to worry about their reliability, robustness, and safety?

Do we need to provide oversight or monitoring of their operation?

How do we guarantee that their behavior is consistent with social norms and human values?

How do we determine when an AI has made the "wrong" decision? Who is liable for that decision?

How should we test them?

For which applications should we use them?

Who benefits from them with regard to standard of living, distribution and quality of work, and other social and economic factors?

Rather than discussing these questions abstractly, one can discuss them using concrete examples. For example: under which conditions, if any, should AI systems be used as part of weapons? Under which conditions, if any, should AI systems be used to care for the handicapped, elderly, or children? Should they be allowed under any conditions to pretend to be human (UK Engineering and Physical Sciences Research Council 2011, Walsh 2016)?

Case Studies

Choices for case studies include anecdotes constructed to illustrate ethical tensions, or actual events (for example, in the form of news stories), or science fiction movies and stories.

News headlines can be used to illuminate ethical issues that are current, visible, and potentially affect the students directly in their daily lives. An example is "Man killed in gruesome Tesla autopilot crash was saved by his car's software weeks earlier" by the *Register* (Thomson 2016), or "Microsoft's racist chatbot returns with drug-smoking Twitter meltdown," by *The Guardian* (Gibbs 2016).

Science fiction stories and movies can also be used to illuminate ethical issues. They are a good source for case studies since they often "stand out in their effort to grasp what is puzzling today seen through the lens of the future. The story lines in sci-fi movies often reveal important philosophical questions regarding moral agency and patiency, consciousness, identity, social relations, and privacy to mention just a few" (Gerdes 2014). Fictional examples can often be more effective than historical or current events, because they explore ethical issues in a context that students often find interesting and that is independent of current political or economic considerations. As Nussbaum puts it, a work of fiction "frequently places us in a position that is both like and unlike the position we occupy in life; like, in that we are emotionally involved with the characters, active with them, and aware of our incompleteness; unlike, in that we are free of the sources of distortion that frequently impede our real-life deliberations" (Nussbaum 1990).

Science fiction movies and stories also allow one to discuss not only ethical issues raised by current AI technology but also ethical issues raised by futuristic AI technology, some of which the students might face later in their careers. One such question, for example, is whether we should treat AI systems like humans or machines in the perhaps unlikely event that the technological singularity happens and AI systems develop broadly intelligent and humanlike behavior. Movies such as *Robot & Frank, Ex Machina,* and *Terminator 2* can be used to discuss questions about the responsibilities of AI systems, the ways in which relationships with AI systems affect our experience of the world (using, for example, Turkle [2012]) to guide the discussion), and who is responsible for solving the ethical challenges that AI systems encounter (using, for example, Bryson, [2016]) to guide the discussion). The creation of the robot in *Ex Machina* can be studied through utilitarianism or virtue ethics.

Teaching Resources

The third edition of the textbook by Stuart Russell and Peter Norvig (2009) gives a brief overview on the ethics and risks of developing AI systems (section 26.3). A small number of courses on AI ethics have been taught, such as by Jerry Kaplan at Stanford University (CS122: Artificial Intelligence — Philosophy, Ethics, and Impact) and by Judy Goldsmith at the University of Kentucky (CS 585: Science Fiction and Computer Ethics). Other examples can be found in the literature (Bates et al. 2012; Bates et al. 2014; Burton, Goldsmith, and Mattei 2015, 2016a). Burton, Goldsmith, and Mattei are currently working on a textbook for their course and have already provided a sample analysis (Burton, Goldsmith, and Mattei 2016b) of E. M. Forster's *The Machine Stops* (Forster 1909). A number of workshops have recently been held on the topic as well, such as the First Workshop on Artificial Intelligence and Ethics at AAAI 2015, the Second Workshop on Artificial Intelligence, Ethics,

and Society at AAAI 2016, and the Workshop on Ethics for Artificial Intelligence at IJCAI 2016. Teaching resources on robot ethics are also relevant for AI ethics. For example, Illah Nourbakhs**h** created an open course website for teaching robot ethics[15] that contains teaching resources to teach a lecture or a whole course on the topic. Several books exist on the topic of machine ethics or robot ethics (Wallach and Allen 2008; Capurro and Nagenborg 2009; Anderson and Anderson 2011; Gunkel 2012; Lin, Abney, and Bekey 2014; Trappl 2015). Case studies can be found at the onlineethics website.[16]

Conclusion

We have provided a case study from the movie *Robot & Frank* as a template for use as is, or as inspiration for discussion of other movies. This case study is not intended to be a complete catalogue of ethical issues or cases, but should function as inspiration and guidance for faculty wanting to devote a few classes to some of the societal implications of the work we do.

Our position is that we as educators have a responsibility to train students to recognize the larger ethical issues and responsibilities that their work as technologists may encounter, and that using science fiction as a foundation for this achieves better student learning, retention, and understanding. To this end some of us have, in the last several years, published work on our course, Science Fiction and Computer Ethics (Bates et al. 2012; Bates et al. 2014; Burton, Goldsmith, and Mattei 2015, 2016b, 2016a). This course has been popular with students, as has Goldsmith and Mattei's previous work running an undergraduate AI course that uses science fiction to engage students about research (Goldsmith and Mattei 2011, Goldsmith and Mattei 2014).

Acknowledgments

Emanuelle Burton and Judy Goldsmith are supported by the National Science Foundation under Grant No. 1646887. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. Research by Benjamin Kuipers at the University of Michigan Intelligent Robotics Lab is supported in part by grants from the National Science Foundation (IIS-1111494 and IIS-1421168). Research by Sven Koenig at the University of Southern California is supported by NSF under grant numbers 1409987 and 1319966.

Some research by Nicholas Mattei was performed while he was employed by Data61, CSIRO (formerly NICTA), and UNSW Australia. Data61, CSIRO (formerly NICTA) is funded by the Australian Government through the Department of Communications and the Australian Research Council (ARC) through the ICT Centre of Excellence Program.

Toby Walsh is supported by the ARC, the ERC, and AOARD.

Notes

1. We create artifacts that take multiple forms including intelligent computing systems and robots. In this article we use the term *AI* and *AIs* to refer to any artificial, autonomous decision maker.

2. See, for example, futureoflife.org/open-letter-autonomous-weapons.

3. futureoflife.org.

4. www.stopkillerrobots.org.

5. webcast.amps.ms.mit.edu/fall2014/AeroAstro/index-Fri-PM.html.

6. www.businessinsider.com/bill-gates-artificial-intelligence-2015-1.

7. www.youtube.com/watch?v=dQOo3Mg4D5A.

8. www.fhi.ox.ac.uk.

9. intelligence.org.

10. An anonymous reviewer suggested that we can summarize Asimov's three laws as decreasing priorities of human-preservation, human-obedience, and robot-self-preservation; the 0th law would be humanity-preservation.

11. arxiv.org/abs/1701.07769.

12. Clip available at youtu.be/eQxUW4B622E.

13. Clip available at youtu.be/3yXwPfvvIt4.

14. Clip available at youtu.be/xlpeRIG18TA.

15. See www.sites.google.com/site/ethicsandrobotics.

16. www.onlineethics.org.

References

Anderson, M., and Anderson, S. L. 2011. *Machine Ethics*. New York: Cambridge University Press. doi.org/10.1017/CBO9780511978036

Annas, J. 2006. Virtue Ethics. In *The Oxford Handbook of Ethical Theory,* ed. D. Copp. Oxford, UK: Oxford University Press.

Anscombe, G. E. M. 2005. Modern Moral Philosophy. In *Human Life, Action and Ethics: Essays by G. E. M. Anscombe,* ed. M. Geach and L. Gormally. Andrews Studies in Philosophy and Public Affairs. Boston: Academic Press.

Aristotle. 1999. *Nichomachean Ethics*, T. Irwin, translator. Indianapolis, IN: Hackett Publishing.

Arkin, R. C. 2009. *Governing Lethal Behavior in Autonomous Robots*. Boca Raton, FL: CRC Press. doi.org/10.1201/9781420085952

Asimov, I. 1950. *I, Robot*. New York: Gnome Press.

Bates, R.; Goldsmith, J.; Berne, R.; Summet, V.; and Veilleux, N. 2012. Science Fiction in Computer Science Education. In *Proceedings of the 43rd ACM Technical Symposium on Computer Science Education*, 161–162. New York: Association for Computing Machinery. doi.org/10.1145/2157136.2157184

Bates, R.; Goldsmith, J.; Summet, V.; and Veilleux, N. 2014. Using Science Fiction in CS Courses. In *Proceedings of the 45th ACM Technical Symposium on Computer Science Education*, 736–737. New York: Association for Computing Machinery.

Bonnefon, J.-F.; Shariff, A.; and Rahwan, I. 2016. The Social Dilemma of Autonomous Vehicles. Unpublished paper deposited in arXiv:1510.03346, Submitted on 12 Oct 2015

(v1), last revised 4 July 2016. Ithaca, NY: Cornell University Libraries.

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

Brynjolfsson, E., and McAfee, A. 2014. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. Boston: W. W. Norton.

Bryson, J. J. 2010. Robots Should Be Slaves. In *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, 63–74. Philadelphia, PA: John Benjamins Publishing Company. doi.org/10.1075/nlp.8.11bry

Bryson, J. J. 2016. Patiency Is Not a Virtue: AI and the Design of Ethical Systems. In *Ethical and Moral Considerations in Non-Human Agents: Papers from the 2016 AAAI Spring Symposium*. Technical Report SS-16-04. Palo Alto, CA: AAAI Press.

Burton, E.; Goldsmith, J.; and Mattei, N. 2015. Teaching AI Ethics Using Science Fiction. In *Artificial Intelligence and Ethics: Papers from the 2015 AAAI Workshop*, 33–37. Palo Alto, CA: AAAI Press.

Burton, E.; Goldsmith, J.; and Mattei, N. 2016a. Using SF to Teach Ethics. Paper presented at the 74th World Science Fiction Convention (Worldcon) Workshop, academic track, Kansas City, MO, 17–21 August.

Burton, E.; Goldsmith, J.; and Mattei, N. 2016b. Using "The Machine Stops" for Teaching Ethics in Artificial Intelligence and Computer Science. In *The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence*. AI, Ethics, and Society: Technical Report WS-16-02, 8-–88. Palo Alto, CA: AAAI Press

Capurro, R., and Nagenborg, M., eds. 2009. *Ethics and Robotics*. Amsterdam, The Netherlands: IOS Press.

Conitzer, V. 2016. Artificial Intelligence: Where's the Philosophical Scrutiny? *Prospect Magazine,* May 4, 2016.

Copp, D. 2005. *The Oxford Handbook of Ethical Theory*. Oxford, UK: Oxford University Press. doi.org/10.1093/0195147790.001.0001

Economist. 2016. The Impact on Jobs: Automation and Anxiety: Will Smarter Machines Cause Mass Unemployment? *The Economist*, June 25, 1016.

Emerging Technology from the arXiv 2015. Why Self-Driving Cars Must Be Programmed To Kill. *MIT Technology Review*, October 22, 2015.

Ford, M. 2015. *Rise of the Robots: Technology and the Threat of a Jobless Future*. New York: Basic Books.

Forster, E. M. 1909. *The Machine Stops*. New York: Simon & Schuster Start Classics.

Frankena, W. K. 1963. *Ethics*. Engelwood Cliffs, NJ: Prentice-Hall.

Gerdes, A. 2014. IT-Ethical Issues in Sci-Fi Film Within the Timeline of the Ethicomp Conference Series. *Journal of Information, Communication and Ethics in Society* 13: 314–325. doi.org/10.1108/JICES-10-2014-0048

Gibbs, S. 2016. Microsoft's Racist Chatbot Returns with Drug-Smoking Twitter Meltdown. *The Guardian,* Wednesday, 30 March 2016.

Goldsmith, J., and Mattei, N. 2011. Science Fiction as an Introduction to AI Research. In *Second AAAI Symposium on Educational Advances in Artificial Intelligence*. Palo Alto, CA: AAAI Press.

Goldsmith, J., and Mattei, N. 2014. Fiction as an Introduction to Computer Science Research. *ACM Transactions on Computing Education* 14(1): 4. doi.org/10.1145/2576873

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

Han, P. 2015. Towards a Superintelligent Notion of the Good: Metaethical Considerations on Rationality and the Good, with the Singularity in Mind. PhD thesis, The Divinity School, The University of Chicago, Chicago, IL.

Johnson, D. G. 2009. *Computer Ethics*. Engelwood Cliffs, NJ: Prentice Hall Press.

Kirilenko, A. A.; Kyle, A. S.; Samadi, M.; and Tuzun, T. 2015. *The Flash Crash: The Impact of High Frequency Trading on an Electronic Market*. SSRN 1686004 Preprint. Amsterdam: Elsevier.

LaFollette, H., and Persson, I. 2013. *Blackwell Guide to Ethical Theory*. London: Wiley-Blackwell.

Lin, P; Abney, K.; and Bekey, G. A., eds. 2014. *Robot Ethics*. Cambridge, MA: MIT Press.

MacIntyre, A. 2007. *After Virtue: A Study in Moral Theory*. South Bend, IN: University of Notre Dame Press.

Maschler, M.; Solan, E.; and Zamir, S. 2013. *Game Theory*. Cambridge, UK: Cambridge University Press. doi.org/10.1017/CBO9780511794216

McNaughton, D., and Rawling, P. 2006. Deontology. In *The Oxford Handbook of Ethical Theory,* ed. D. Copp. Oxford, UK: Oxford University Press. doi.org/10.1002/9780470510544.ch9

Nussbaum, M. 1990. *Love's Knowledge: Essays on Philosophy and Literature*. Oxford, UK: Oxford University Press.

Patila, I.; Cogonia, C.; Zangrandob, N.; Chittarob, L.; and Silania, G. 2014. Affective Basis of Judgment-Behavior Discrepancy in Virtual Experiences of Moral Dilemmas. *Social Neuroscience* 9(1): 94–107. doi.org/10.1080/17470919.2013.870091

Piketty, T. 2014. *Capital in the Twenty-First Century*. Cambridge, MA: Harvard University Press. doi.org/10.4159/9780674369542

Ripstein, A. 1989. Review of Russell Hardin, "Morality Within the Limits of Reason." *Canadian Journal of Political Science / Revue Canadienne de Science Politique* 22(3): 685–686. doi.org/10.1017/S0008423900011458

Russell, S.; Dewey, D.; and Tegmark, M. 2015. Research Priorities for Robust and Beneficial Artificial Intelligence. *AI Magazine* 36(4): 105–114.

Russell, Stuart, and Norvig, Peter 2009. *Artificial Intelligence: A Modern Approach*. 3rd edition. London: Pearson.

Schumacher, E. F. 1979. *Good Work*. New York: Harper & Row.

Thomson, I. 2016. Man Killed in Gruesome Tesla Autopilot Crash Was Saved by His Car's Software Weeks Earlier. *The Register*, 30 June 2016.

Trappl, R., ed. 2015. *A Construction Manual for Robots' Ethical Systems*. Berlin: Springer.

Turkle, S. 2012. Alone Together: Why We Expect More from Technology and Less from Each Other. *New York:* Basic Books.

UK Engineering and Physical Sciences Research Council. 2011. *EU Principles of Robotics*. London: Engineering and Physical Sciences Research Council. (www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics/. Downloaded 3-25-2016).