Teaching AI Ethics Using Science Fiction

Emanuelle Burton
Centre College
Danville, Kentucky, USA
emanuelle.burton@centre.edu

Judy Goldsmith
University of Kentucky
Lexington, Kentucky, USA
goldsmit@cs.uky.edu

Nicholas Mattei
NICTA and UNSW
Sydney, Australia
nicholas.mattei@nicta.com.au

Abstract

The cultural and political implications of modern AI research are not some far off concern, they are things that affect the world in the here and now. From advanced control systems with advanced visualizations and image processing techniques that drive the machines of the modern military to the slow creep of a mechanized workforce, ethical questions surround us. Part of dealing with these ethical questions is not just speculating on what could be but teaching our students how to engage with these ethical questions. We explore the use of science fiction as an appropriate tool to enable AI researchers to help engage students and the public on the current state and potential impacts of AI.

Introduction

We can look at the movie WARGAMES (Lasker and Parkes 1983) and consider the ethical conundrums of letting a smart system handle our military capabilities, or we can extrapolate to the possible future so grimly portrayed in the Terminator movies (Cameron and Wisher 1997) and so many others. We can speculate philosophically about AIs and the extinction of humans (Andersen 2013) or the necessity of "The Three Laws" in *iRobot* (Asimov 1950). But we have real, present ethics violations and challenges arising from current AI techniques and implementations, in the form of systematic decreases in privacy; increasing reliance on AI for our safety, and the ongoing job losses due to mechanization and automatic control of work processes.

Although some ethicists engage in comparative study whose purpose is largely descriptive, nearly all work in ethics – both academically and in the wider world – is, by contrast, normative: that is, it argues how the world should be understood, and how people ought to act. Most approaches to ethics adopt one of three basic postures. Virtue ethics (also known as teleological ethics), grounded in classical philosophy and outlined most clearly in Aristotle, is organized around developing habits and dispositions that are conducive to acting virtuously and, by extension, to flourishing as an individual (Annas 2006). Other approaches are comparatively recent; deontological (or law-based) ethics, developed by Immanuel Kant, argues that means, rather than ends, are similar, and recent theorists have argued that virtue ethics is best seen as part of successful deontology (McNaughton and Rawling 2006). A still more recent approach, utilitarian ethics was developed by Jeremy Bentham and John Stuart Mill in the late 18th to mid-19th century, and seeks the greatest good for the greatest number – or, in William K. Frankena's formulation, "the greatest possible balance of good over evil (Frankena 1963)." While all three schools have proponents among philosophers, theologians, and other scholars who work in ethics, broader discourse tends to presume the merits of a utilitarian approach. This shared assumption, however, does little to provide a shared baseline for conversation about ethics; utilitarianism is theoretically insubstantive, and offers no critical resources for filling in the particulars of "the greatest good for the greatest number."

Given that our work as AI professionals comes up against ethical issues on such a regular basis, it is incumbent on us to educate our students about the ethical issues that are arising and are likely to arise in our, and their, careers. The burgeoning possibilities for AI put technologists (and possibly their creations) in situations that call not only for technical decisions but for ethical ones. The precise criteria of moral judgment varies according to different schools of ethical theory, but most readers can easily recognize that the powers at stake in these new technologies are ones that render both individuals and societies vulnerable, and that students being trained in the field need to be trained to recognize the issues at stake and to understand the responsibility that their work as technologists will entail.

Among researchers in the AI community, there are not only multiple sets of values, but different approaches to the theory of value. Different sets of values express themselves in obvious ways: many of us work for military entities, are funded by military entities, or do work with military applications, whether or not those are the intended applications. Some of us are pacifists. Some of us believe that an ideal society is one where individuals are free of the requirement to work, while others of us believe that work creates society and a social contract. However, we also seem to have different approaches to the the theory of value, i.e., what we have a responsibility to do. While some hold that our work as technologists is objective, that we create without affecting, others believe that it is part of our jobs to advocate for social and political change, based on our understanding of technology. We are not attempting to unify the field around a particular value system. Rather, we are advocating that we must come to terms with a basic fact: our work does not take place in an ethical vacuum, and whatever principles we establish for AI will have ethical implications. We can, and we must, come to terms with this basic fact, even if we continue to differ about the appropriate value systems on which to build AI systems. The goal of teaching ethics is not to impose a value system on our students, but to make them aware of the social ramifications of their work, and the possibility that research, development, and implementation can be carried out in a variety of ways. We want our students to make informed, thoughtful, and ethical choices.

Contemporary Ethical Questions in AI

Surveillance and eavesdropping have been the common currency of state intelligence agencies as long as there has been written history. Indeed, Sun Tzu's *The Art of War* contains sections detailing the necessity of gathering intelligence about both allies and enemies. However, in recent years, the adoption of advanced techniques within AI, mathematics, and surrounding communities have made this practice common, sophisticated, and available at a scale never seen before (Greenwald 2014). While there are clear cut value-based arguments for and against this practice, the reality is here: almost everything done online is being monitored. Where the normal apparatus of state surveillance used to employ massive resources in terms of person power and time, now a small group of programmers can monitor entire networks of individuals that may be physically located anywhere on the globe.

Big data techniques involve not only massively parallel search and association learning, but network analysis. The ethical issues associated with network analysis at a national or global scale certainly involve the value and rights to privacy. They also include safety, in that such analysis is used to broaden targets (of government or terrorist actions) from individuals to those with whom they associate. Thus, a backchannel interaction, such as physical or social proximity, with a targeted individual or organization, can put us in danger, whether or not we are "legitimately" part of that target. This includes civilian deaths from drone strikes, or arrests of suspected terrorists and their social circles. Much like in the movie *Brazil* (Gilliam, Stoppard, and McKeown 1985), when Archibald Buttle is mistakenly executed instead of Archibald Tuttle, drawing the main character Sam Lowery, by association, into a downward spiral ultimately resulting in his torture and death, the threat of associating with the wrong people has become a much more pressing concern.

Ethical issues in computer science pop up in many more, seemingly innocuous, areas. The race to perfect the driverless car is born from sobering statistics: over 33,000 people a year die in traffic accidents in the US alone. Sebastian Thrun explains in his TED Talk (Thrun 2011) that *preventing* these deaths was a key motivator in developing driverless car technology. Regardless, given the stated goal of the Stanford (and Google) driverless car project is to show "... how ridiculous it was that humans were driving cars," the program has ethical considerations in other areas. According to the US Bureau of Labor Statistics[1] there are currently 300,000 cab drivers, 1.7 million truck drivers, 700,000 bus drivers, 1.2 million delivery truck drivers, and 650,000 material movers, roughly 3% of the total employed workforce of 146 million people. We are actively trying to eliminate these jobs from society, with no meaningful replacement. This "end of work" discussion is not a new one, though the ongoing "jobless recovery" in the first world (Desliver 2013), among other arguments (Anonymous 2010), indicates that much of the *intelligent* mechanization being developed by the AI community, from algorithms to write reviews and sports stories (Birnbaum 2013) to smart machine that do everything from spraying crops, to acting as cabs, to moving material around mines, is eliminating jobs at a scale previously unseen (Grey 2014; Anonymous 2014) while not significantly impacting productivity. While may of these pursuits are undertaken with good intentions, there are ethical considerations that we, as researchers, need think about and discuss. We, as leaders, need to engage with the broader community about these issues.

Using Science Fiction to Teach

In previous work we have outlined an undergraduate computer science course that centers on using science fiction to engage students in a first course on research methods and ethical issues in a non-threatening way (Goldsmith and Mattei 2011; 2014). We have even gone so far as to run a course on "Science Fiction and Computer Ethics" (Mihail, Rubin, and Goldsmith 2014), to great success. We are not alone in this endeavor or this opinion as others have cited using science fiction as a gateway as it, "[...] often removes the intellectual and emotional resistance some students might at first feel towards the subject of ethics (Pease 2009)." Pease reports that using science fiction to teach *practical ethics* engages students from a variety of majors including computer science and engineering. Additionally, Bates uses science fiction as motivation to talk about AI in a general education (nonmajors) course, and as an entry point for talking about ethics (Bates 2011). Courses in other fields use science fiction for motivation. Bowring and Tambe list two, including an intriguing case that used science fiction as a tool for teaching children to think about the future (Dils 1987).

The use of AI as a hook for CS participation and discussion is not a new one. AAAI has had a symposium on "Using AI to motivate greater participation in Computer Science" (Sahami 2008). Using Artificial Intelligence to inspire young researchers and scientists is extremely popular (Sawyer 2002; Watson 2003). Additionally, AAAI maintains a list of places that AI is used in society, providing even more ideas for application areas ripe for discussion (AAAI 2011). SIGCSE has had a panel (Bates et al. 2012) and a birds of a feather session (Bates et al. 2014) on using science fiction to teach computer science; this year's SIGCSE will have a birds of a feather session dedicated to discussion of three CS-related science fiction novels (Ready Player One, I, Robot, and Bellwether).

[1] http://www.bls.gov/home.htm

In computer science, as with many high consensus fields, there is a tendency to teach from authority and not encourage discussion and dissent (Colbeck 1998). While many feel that this is a necessary evil in order to establish a firm foundation of core knowledge within students, there is no doubt that it reinforces "absolute truth" views of knowledge (Haworth and Conrad 1995). Students may graduate with gaps in their ability to think and reason about situations which involve ethical decisions, which often have more than one "correct" answer.

Engaging multiplicity and other forms of critical thinking through exposure to AI and ethics research and writing will provide our students with examples of thinking that move beyond dualism and other didactic modes of reasoning (Davis 2009; Perry 1980). Using reading and writing to both expand students' communication skills and to gain understanding and insight into the topics that the student is currently learning (Walker 1998; Hoffman, Dansdill, and Herscovici 2006; Bean 2011; McLeod and Soven 1992) is a popular idea in curicculum design. Without addressing and fostering our students' ability to think critically, we may forever leave them unable to judge multiple solutions in a principled and discerning way.

Fiction

Literature has long been held to play a role in shaping the beliefs of its readers; ever since Socrates banned the poets from his ideal city of Kallipolis in Book X of Plato's Republic, on account of the dangers their work posed, philosophers and religious thinkers have gloried in and despaired of the power of literature to make or break a reader's convictions about the world. In recent years, many ethical scholars have even argued that literature is superior to philosophy in its ability to represent and address the ethical conundra of human experience. Martha Nussbaum, one of the preeminent exponents of this position, writes,

> Reading [fiction] frequently places us in a position that is both like and unlike the positon 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. Since the story is not ours, we do not get caught up in the vulgar heat of our personal jealousies or angers or the sometimes blinding violence of our loves. (Nussbaum 1990)

Although the literary and philsoophical establishment has not historically taken science fiction seriously as a venue for ethical thinking, this fact reflects longstanding biases in the field rather than the merits or possibilities of science fiction itself. By reframing recognizable human situations and problems in terms of unfamiliar settings and technology, science fiction (like fantasy fiction) can be an ideal medium for raising and exploring ethical concerns. By presenting a familiar problem (such as conflicts between different social groups or the invasion of privacy) in unfamiliar terms and settings, a work of science fiction can limit a reader's ability to identify transparently with any one aspect or position. This isolation can create a safer imaginative space for the reader to explore a wider range of possible sympathies. Furthermore, the strangeness of the setting or context can make visible the alarming or problematic aspects of a given situation that have become invisible in the mundane world because they have come to be regarded as ordinary or inevitable.

Why *Science* Fiction

"Science fiction is a genre of fiction dealing with imaginative content such as futuristic settings, futuristic science and technology, space travel, time travel, faster than light travel, parallel universes, and extraterrestrial life. (Wikipedia 2014)"

Arguably, any fiction that raises significant technical or ethical questions about computers, robots, or computer science *is* science fiction. We focus on science fiction for two reasons. One, the use of futuristic or alien settings allows students to detach from political preconceptions and experience the dilemmas of plot and characters as something fresh.

Two, it has so far proved popular with the students. They have perceived that the course would be a chance to get credit for something they enjoy but have not found time to do while in college/graduate school: read and watch science fiction.

There has been discussion amongst colleagues who teach similar courses about whether the use of science fiction has been a turn-off for women. We have not found it to be. It is important that any course built on science fiction make sure that some of the works used pass the Bechdel test[2] (Bechdel 1986).

Case Studies

Consider Melissa Scott's The Jazz (Scott 2001). The novel raises many ethical issues of parental responsibility for underage hacking, and more generally of society's obligations to vulnerable users. The plot is driven by a teenager hacking into a movie studio's system and stealing code which the studio executive is desperate to recover. Ultimately, we learn the meaning of the stolen software. It predicts audience reactions to movie elements, and drives the "creative" process at that studio. This raises questions about the appropriate role of AI in the arts.

Neal Stephenson has raised many interesting ethical issues, such as the use of online games to launder money internationally in *REAMDE* (Stephenson 2011). In *Diamond Age*, he looks at computer mediated and computer-controlled education, in the forms of the Young Ladies' Primer (Stephenson 2000). While most educational effort in AI these days is on smart tutors and automatic assignment grading, we can extrapolate a trend toward AI teachers. Stephenson presents two forms of that. In one, the Primer produces scripts and scenarios that a human actress modifies. In the other, goals have been defined, but the actual educational process is (perhaps) not monitored by any human. The fact that this produces an army of children, brainwashed

[2] The Bechdel test asks if a work: (1) has at least two women in it, (2) who talk to each other, (3) about something besides a man.

to obey one particular child, is presented as a good thing, or at least useful to the plot, but it certainly raises red flags about the ethics of AI-driven education.

Novels such as Scott Westerfeld's *Extras* (Westerfeld 2007) and Gary Shteyngart's *Super Sad True Love Story* (Shteyngart 2011) make vivid the role that computer-mediated reputations can play in possible future societies. Mary Doria Russell's Catholic science fiction novel, *The Sparrow* (Russell 1997), raises important issues about AI practitioners taking over human expertise and automating professions. David Eggers' *The Circle* (Eggers 2013) rather melodramatically considers the possible consequences of ubiquitous monitoring.

Fiction doesn't have to be in the form of books, nor does it need to be recent to be effective. Charlie Chaplin's 1936 movie, *Modern Times* (Chaplin 1936), was a strong political statement against industrialization and mechanization of production. In it, Chaplin's character, the little tramp, gains and loses several jobs, suffers several arrests, and poverty. It is most memorable for scenes of Chaplin struggling to keep up with an assembly line, and being the guinea pig for an entirely unsuccessful "feeding machine". His first arrest happens at a Communist demonstration that he falls into by accident.

The movie has proved popular with the students, and serves as an entry into the discussion of job loss due to technology. It is a surprisingly easy arc from the frenetic assembly line of the movie to robotic assembly lines in today's factories, and from there job loss due to automated phone directories, self-driving cars, etc.

Results

In the Spring of 2013 the course *Interim Topics in Computer Science: Science Fiction and Computer Ethics*, was offered at the University of Kentucky. The course had an enrollment of 29 and 22 of those students filled out course evaluations at the end of the semester. While this is a small sample size and all course evaluations should be taken with a grain of salt, we report feedback from the students here. The course evaluations at the University of Kentucky are rated on a 4 point Likert scale with 1 being *Strongly Disagree* and a 4 being *Strongly Agree*. The course review forms were overwhelmingly positive with the only 5 responses below *Agree* across all 22 questions for all 22 students. The the two summary questions for the course were both extremely positive with the mean student response to the "Overall Value of the Course" of 3.9/4.0 (20 Strongly Agree (SA), 2 Agree (A) ratings) and to the "Overall Value Quality of the Teaching" of 4.0/4.0 (21 SA and 1 A).

Many students found the class enjoyable and echoed that it should be considered as a required course for all students in computer science and, more broadly, in engineering:

> *... this class has been interesting and informative. I think every engineer/CS etc. [student] should be required to take an ethics course.*

Another student supported this idea:

> *I think this would be an excellent course to offer as part of the regular curriculum. The content was very engaging and always enjoyable. [...] After having completed the course, I feel this might even be a useful required course for the CS degree.*

The course evaluations also includes questions which relate to specific learning outcomes. One reason for teaching ethics is to engage the students and challenge them to look at their work and its purpose with a critical eye. The course evaluation question, "This course strengthen my ability to analyze and evaluate information," had a mean student response of 3.6/4.0 (18 SA and 6 A). This learning outcome was also reinforced in several student comments including:

> *Very interesting topic that made you think about a lot of things you may not have thought [about] when it comes to ethics.*

We not only want the students to think critically but also to respect the pluralism of viewpoints that exist. On the course evaluation question "I learned to respect different viewpoints different from my own," the students responded with a mean of 3.7/4.0 (15 SA, 7 A). More interestingly, several students specifically mentioned this outcome in their comments as being a positive and useful outcome of the course:

> *I think this course should be a requirement for undergrad students, because it helps people understand the ramifications of their coding, more than the day or two in CS100*[3]*. It also helps people respect another's opinion, and a lot of programmers are not respectful of others, and hold the "I can do it better on my own" opinion.*

A final learning outcome we were pleased to see – though it is likely more of a result of the format of the course than the topic – was the responses to the question, "Encouraged class participation," where the students responded with a mean of 4.0/4.0 (21 (SA) and 1 (A). Several student comments were received like the one below, which are extremely encouraging:

> *This course helped me to feel more comfortable speaking with students and presenting my opinion. I learned to understand and analyze different viewpoints and to think before responding. I would not increase the class size. Larger classes would make it more difficult for students who are introverts to feel comfortable.*

These three learning outcomes, taken together, imply that we achieved our goal: we wanted students to engage with a plurality of viewpoints; analyze and think critically about these viewpoints; and engage in discussion with their peers in a constructive and respectful way. While the sample size is small and we would want to see results from other instructors and other institutions, we feel that this represents a strong start.

Conclusions

In short, we hold that current AI practices and research raise many ethical issues. We are concerned that the focus of AI

[3] CS100 is the Introduction to the Computer Science Professions course at the University of Kentucky

ethics discussions are on future technology rather than the very present dangers of current and very-near-future tech, and we advocate for strong ethics training as part of CS education for students and practitioners. In the academic setting, we suggest that science fiction may be used as an effective tool to immerse students in case studies where they can debate the appropriateness of decisions made, and consider possible consequences of those decisions in a less threatening, though still imaginable, fictional world.

References

AAAI. 2011. AI topics: Science fiction: Views of the future involving AI. http://www.aaai.org/AITopics/pmwiki/pmwiki.php/AITopics/ScienceFiction.

Andersen, R. 2013. Omens. *Aeon*. http://aeon.co/magazine/philosophy/ross-andersen-human-extinction/.

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

Anonymous. 2010. Automatic reaction. *The Economist*. 9 Sep.

Anonymous. 2014. The onrushing wave. *The Economist*. 18 Jan.

Asimov, I. 1950. *I, Robot*. 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. ACM.

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. ACM.

Bates, R. A. 2011. AC 2011-1669: AI & SciFi: Teaching writing, history, technology, literature and ethics. In *ASEE Annual Conference & Exposition*.

Bean, J. C. 2011. *Engaging ideas: The professor's guide to integrating writing, critical thinking, and active learning in the classroom*. John Wiley & Sons.

Bechdel, A. 1986. *Dykes to Watch Out For*. Firebrand Press.

Birnbaum, L. 2013. Telling stories at internet scale. Joint AAAI/IAAI 2013 Invited Talk.

Cameron, J., and Wisher, W. 1997. Terminator 2: Judgement day. Director: James Cameron.

Chaplin, C. 1936. Modern times. Director, Charles Chaplin.

Colbeck, C. L. 1998. Merging in a seamless blend: How faculty integrate teaching and research. *The Journal of Higher Education* 69(6):647–671.

Davis, B. G. 2009. *Tools for Teaching*. Jossey-Bass, second edition.

Desliver, D. 2013. At 42 months and counting, current job recovery is slowest since Truman was president. *The Pew Research Center*. 25 Sept.

Dils, L. S. 1987. Science fiction and the future, Yale-New Haven Teachers Institute. http://www.yale.edu/ynhti/curriculum/units/1987/2/87.02.04.x.html.

Eggers, D. 2013. *The Circle*. Knopf.

Frankena, W. K. 1963. *Ethics*. Prentice-Hall.

Gilliam, T.; Stoppard, T.; and McKeown, C. 1985. Brazil. Director: T. Gilliam.

Goldsmith, J., and Mattei, N. 2011. Science fiction as an introduction to ai research. In *Proc. 2nd AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI)*.

Goldsmith, J., and Mattei, N. 2014. Fiction as an introduction to computer science research. *ACM Transactions on Computer Science Education* 14(1):4:1–4:14.

Greenwald, G. 2014. *No Place to Hide: Edward Snowden, the NSA, and the U.S. Surveillance State*. Metropolitan Books.

Grey, C. 2014. Humans need not apply. https://www.youtube.com/watch?v=7Pq-S557XQU.

Haworth, J. G., and Conrad, C. F. 1995. Curricular transformations: Traditional and emerging voices in the academy. In Haworth, J. G., and Conrad, C. F., eds., *Revisioning Curriculum in Higher Education*. Simon & Schuster Custom Publishing. 191 – 202.

Hoffman, M. E.; Dansdill, T.; and Herscovici, D. S. 2006. Bridging writing to learn and writing in the discipline in computer science education. In *Proceedings of the 37th ACM Technical Symposium on Computer Science Education (SIGCSE 2006)*, 117–121.

Lasker, L., and Parkes, W. F. 1983. WARGAMES. Directed by John Badham.

McLeod, S. H., and Soven, M. 1992. *Writing across the curriculum*. Sage Publications.

McNaughton, D., and Rawling, P. 2006. Deontology. In Copp, D., ed., *The Oxford Handbook of Ethical Theory*. Oxford University Press.

Mihail, P.; Rubin, B.; and Goldsmith, J. 2014. Online discussions: Improving education in CS? In *Proceedings of the 45th ACM Technical Symposium on Computer Science Education (SIGCSE 2014)*.

Nussbaum, M. 1990. *Loves Knowledge: Essays on Philosophy and Literature*. Oxford University Press.

Pease, A. 2009. Teaching ethics with science fiction: A case study syllabus. *Teaching Ethics: The Journal of the Society for Ethics Across the Curriculum* 9:75–82.

Perry, W. G. 1980. Cognitive and ethical growth: The making of meaning. In Chickering, A. W., and Associates., eds., *The Modern American College*. Josey Bass. 76–109.

Russell, M. D. 1997. *The Sparrow: A Novel*. Ballantine Books.

Sahami, M. 2008. Using AI to motivate greater participation in computer science, AAAI Spring Symposium. Technical Report SSS08. AAAI Spring Symposium.

Sawyer, R. J. 2002. AI and sci-fi: My, oh, my! Keynote Address to The 12th Annual Canadian Conference on Intelligent Systems.

Scott, M. 2001. *The Jazz*. Tor Books.

Shteyngart, G. 2011. *Super Sad True Love Story*. Random House.

Stephenson, N. 2000. *Diamond Age*. Spectra.

Stephenson, N. 2011. *REAMDE*. William Morrow Paperbacks.

Thrun, S. 2011. Google's driverless car. TED 2011.

Walker, H. M. 1998. Writing within the computer science curriculum. *ACM SIGCSE Bulletin* 30(2):24–25.

Watson, I. 2003. The aims of artificial intelligence: A science fiction view. *IEEE Intelligent Systems* 18.

Westerfeld, S. 2007. *Extras*. Simon & Schuster.

Wikipedia. 2014. Science fiction — wikipedia, the free encyclopedia. [Online; accessed 13-October-2014].