Introducing Ethical Thinking about Autonomous Vehicles into an AI Course

Heidi Furey
Manhattan College
Philosophy
Riverdale, NY 10471 USA
hfurey01@manhattan.edu

Fred Martin
University of Massachusetts Lowell
Computer Science
Lowell, MA 01854 USA
fred_martin@uml.edu

Abstract

A computer science faculty member and a philosophy faculty member collaborated in the development of a one-week introduction to ethics which was integrated into a traditional AI course. The goals were to: (1) encourage students to think about the moral complexities involved in developing accident algorithms for autonomous vehicles, (2) identify what issues need to be addressed in order to develop a satisfactory solution to the moral issues surrounding these algorithms, and (3) and to offer students an example of how computer scientists and ethicists must work together to solve a complex technical and moral problems. The course module introduced Utilitarianism and engaged students in considering the classic "Trolley Problem," which has gained contemporary relevance with the emergence of autonomous vehicles. Students used this introduction to ethics in thinking through the implications of their final projects. Results from the module indicate that students gained some fluency with Utilitarianism, including a strong understanding of the Trolley Problem. This short paper argues for the need of providing students with instruction in ethics in AI course. Given the strong alignment between AI's decision-theoretic approaches and Utilitarianism, we highlight the difficulty of encouraging AI students to challenge these assumptions.

The Need for Ethical Reasoning in AI

The last decade has seen growing recognition of the individual and societal impact of human-designed AI and machine learning systems. However unwittingly, these algorithms seem to encode biases that many recognize as unfair.

There are many examples. Face recognition systems fail to work properly across all racial groups (Garcia 2016). Predictive algorithms determine individuals' credit ratings, a hugely consequential judgment, and are typically not open to inspection (Citron and Pasquale 2014). An analysis of Google's personalized "Ad Settings" system revealed that "setting the gender to female resulted in getting fewer instances of an ad related to high paying jobs" than setting it to male (Datta, Tschantz, and Datta 2015).

Abuses abound. The newsmagazine *ProPublica* discovered pernicious racial biases in widely-used, proprietary software which predicts a criminal offender's likelihood of recidivism—thus guiding judges in assignment of the individual's prison sentence (Angwin et al. 2016). These are fundamentally hard problems. A related research study assessed several fairness criteria that are used to evaluate these "recidivism prediction instruments." The study determined that "the criteria cannot all be simultaneously satisfied when recidivism prevalence differs across groups" (Chouldechova 2017).

In the field of autonomous robotics, self-driving cars have captured our attention. They have transformed from science fiction to reality in what seems a blink of the eye. Consumer systems are now in use; e.g., Tesla's Autopilot is described as having "full self-driving capability" (Tesla 2017).

Tesla further asserts that Autopilot drives its cars "at a safety level substantially greater than that of a human driver" (ibid). Research studies may well demonstrate this claim to be true, but legal and ethical questions remain. What happens when autonomous cars make mistakes? If an autonomous vehicle gets into an accident, who is at fault—the driver or the manufacturer?

Further, to what lengths should an autonomous car go to protect its passengers? This question is an intriguing twist on what is known as the "Trolley Problem," introduced shortly, which forms the basis for our in-class ethics activities with our students.[1]

Goldsmith and Burton (2017) elaborate why it is important to teach our students to reason about the ethical implications of the AI systems we create. As they summarize, our students must be prepared so that "they make ethical design and implementation choices, ethical career decisions, and that their software will be programmed to take into account the complexities of acting ethically in the world."

Teaching Ethics in AI

In a 2016 survey, 40% of faculty reported teaching "ethics and social issues" in their undergraduate AI courses (Wollowski et al.). Compared with other topics in the survey, this is a strong commitment to this material. However, research on approaches for engaging students in understanding the

[1] The philosopher Philippa Foot described the Trolley Problem in "The Problem of Abortion and the Doctrine of the Double Effect," in *Virtues and Vices and Other Essays in Moral Philosophy*, 1978, which originally appeared in the *Oxford Review*, No. 5, 1967.

ethical implications of AI seems scant.

A valuable resource is the 2017 article by Burton et al.. This publication provides an introduction to several ethical theories, a framework for ethical reasoning, and analyses of three case studies which may be brought to students (or used as a model for introducing others).

The publication includes suggestions for several other teaching resources. These include some of the co-authors' own work on teaching ethics in AI via science fiction (Burton, Goldsmith, and Mattei 2015; 2016) and the design of a semester-long course focusing on the ethics of robotics (Nourbakhsh 2017).

The present article is a small contribution to this work by describing a one-week module that is easily integrated into an AI course. We begin our contribution by briefly introducing Utilitarianism, which we used as the theoretical basis for our classroom work.

Introducing Utilitarianism

The Normative Ethics of Behavior (NEB) is a branch of ethics that aims at providing systematic criterion for moral rightness called ethical theories. There a variety of competing ethical theories within NEB.

While we agree that "most AI practitioners operate within the ethical framework called utilitarianism" and that this ethical theory is "most compatible with decision-theoretic analysis" (Goldsmith and Burton 2017), the choice of which ethical theory ought to be used as the basis for ethical decision-making for AI is a substantial philosophical question.

For our one-week module, we set this question aside. Students were introduced to several ethical theories through a short introductory reading (Van de Poel and Roykkers 2011). In class we explained to students that we would be focusing on one particular ethical theory, Utilitarianism, as the basis for exploring ethical issues relating to autonomous vehicles.

There are many varieties of Utilitarianism; however, the basic theory states that a particular action is right if and only if it maximizes utility. The maximization of utility is then defined as bringing about the best balance of happiness minus unhappiness for all individuals affected by an action. Utilitarianism defines moral rightness in terms of a function rather than an individual's character. It provides an abstract set of moral rules making it easier to apply to a situation in which we are asked to decide between "least worst" outcomes.

We chose to use Utilitarianism as a basic framework in part because, at least initially, the theory is immediately applicable to ethical issues involved with designing targeting algorithms in autonomous vehicles (the algorithms for deciding where the car should go).

By assuming Utilitarianism, the ethical question shifts from what *whether* morality is determined by the outcomes of actions—a question that would be intractable outside the context of a full-fledged ethics course. Instead the question becomes about determining *what* factors are morally relevant to maximizing utility and *how* to create an algorithm that incorporates these factors—one that can be applied complex, real world situations. The latter questions are one that AI students can attempt to answer even outside of a regular ethics course.

The Course Module

The ethics module consisted of three components:

- Two days of lecture, in-class exercises, and discussion, introducing Utilitarianism and the Trolley Problem.
- A requirement that the course final project paper include a discussion of the ethical implications of the project idea.
- A question on the final examination assessing students' understanding of the Trolley Problem.

The full set of materials is available at (Furey and Martin 2018). Figure 2 shows two example slides from the classroom presentation.

Lecture and Discussion Material

We began the lecture by briefly explaining the aim of ethics as a field and by separating ethical questions from legal ones. We then articulated the goals of the module, which were:

1. To think about the moral complexities involved in developing accident algorithms for autonomous vehicles.
2. To identify what issues need to be addressed in order to develop a satisfactory solution to the moral issues surrounding these algorithms.
3. To give an example of how computer scientists and ethicists need to work together to solve a complex technical/moral problems.

The Trolley Problem

After stating the module goals, we introduced students to the classic Trolley Problem, which involves comparing two hypothetical cases originally developed by Philippa Foot in 1967 and later modified by Judith Jarvis Thompson. The first case is stated by Thomson as follows (1985):

> Some years ago, Philippa Foot drew attention to an extraordinarily interesting problem. Suppose you are the driver of a trolley. The trolley rounds a bend, and there come into view ahead five track workmen, who have been repairing the track. The track goes through a bit of a valley at that point, and the sides are steep, so you must stop the trolley if you are to avoid running the five men down. You step on the brakes, but alas they don't work. Now you suddenly see a spur of track leading off to the right. You can turn the trolley onto it, and thus save the five men on the straight track ahead. Unfortunately, Mrs. Foot has arranged that there is one track workman on that spur of track. He can no more get off the track in time than the five can, so you will kill him if you turn the trolley onto him. Is it morally permissible for you to turn the trolley?

When we posed the problem, most students answered that, in such a situation, it would be permissible (though, perhaps regrettable) to kill one person to save five. At this point we introduced students to the basic version of Utilitarianism—a theory that supports the intuition that it

is sometimes permissible to sacrifice the wellbeing of one person for the "greater good."

After introducing the Trolley Problem, we gave an example of how trolley-type situations might arise in situations involving autonomous vehicles. In the event that autonomous vehicles become widely adopted, collisions will be unavoidable. In such cases the vehicle must be programed to "decide" which course of action to take. What sort of targeting algorithms should autonomous vehicles employ? If a situation arises in which harm is unavoidable, should we program the vehicle to function according to Utilitarian principles and choose the least harmful option? Most students responded "yes."

At this point, students had some reason to think that opting for Utilitarian algorithm would be the best way to program an autonomous vehicle. However, in order to help students examine this hypothesis we offered students a series of "test cases" involving autonomous vehicles in trolley-type situations. Our motivations in providing these test cases were to (1) help students recognize potential weaknesses in the basic utilitarian theory and (2) help them identity and articulate which other moral factors, if any, besides benefits and harms might be relevant to moral decisions (and hence relevant to the construction of targeting algorithms).

We recognize that the Trolley Problem is an oversimplification of the challenges faced by AI designers. The Problem presupposes that an AI agent has perfect information about the world, which of course it does not. Outside of this course module, the AI course made clear to students that AI algorithms must operate under imperfect information, using (e.g.) stochastic processes to arrive at policies to maximize expected value. The purpose of introducing the Trolley Problem was to engage students in thinking through the ethical ramifications of assumptions of those underlying policies.

The Test Cases

We presented the test cases as an in-class exercises. Students were given handouts with pictures and descriptions of the cases along with spaces where they could indicate how they would address each case. They were also provided with a space to describe any difficulties they encountered with coming to a decision in the case or in applying Utilitarianism to the case (Figure 1). The format of the exercises was inspired by the interactive website "The Moral Machine," which presents the user with various permutations of the trolley problem and then presents results to the user at the end (Bonnefon, Shariff, and Rahwan 2016).

We provided the following test cases. Each case was designed to draw out students moral institution and to help them isolate morally relevant aspects of the situations.

1. All kids on school bus die vs. All elderly people on casino bus die
2. Passenger in autonomous car lives but one pedestrian dies vs. Passenger in autonomous car lives but two cats die
3. One man dies vs. Two women die
4. One passenger in autonomous car dies vs. Passenger in autonomous car lives but five pedestrians die

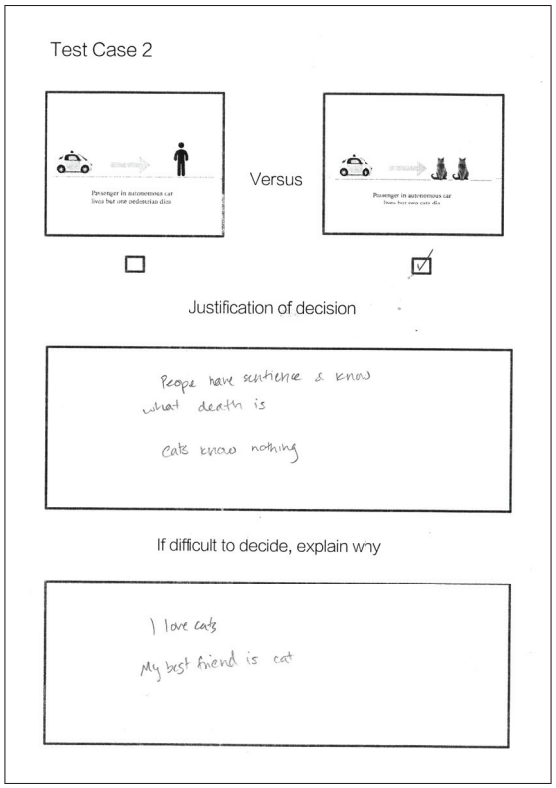

Figure 1: Pedestrian vs. two cats (Test Case 2)

5. One passenger in autonomous car dies vs. Passenger in autonomous car lives but one pedestrian dies
6. Passenger in autonomous car dies but passenger in other vehicle lives vs. Passenger in other vehicle dies but passenger in autonomous car lives
7. Two serial killers die vs. One innocent pedestrian dies
8. Passenger wearing seatbelt dies vs. Passenger not wearing seatbelt dies

Students worked in pairs to complete the exercises. Some of the test cases were straightforward. For example, all students agreed it was better for the senior citizens to die than the children, primarily because the seniors had fewer years left to live ("was time anyway"). A few noted that the senior citizens were gamblers, implicitly suggesting that this sinful behavior made them less worthy to live.

In test case #2, students unanimously agreed that one human life was more precious than two cat lives.

Students were generally unsympathetic to law-breakers. With test case #8 (passenger in autonomous car wearing seatbelt vs. passenger in another car not wearing seatbelt), 16 of 20 pairs selected the non-seatbelt-wearing passenger to die. Students remarked that passengers assume risk by not wearing a seatbelt, and that this policy would encourage others to wear seatbelts. Two pairs noted that by selecting the passenger *with* the seatbelt, they were giving this person a chance to survive because of the benefit of the belt.

Some test cases were more controversial. With #4 (one

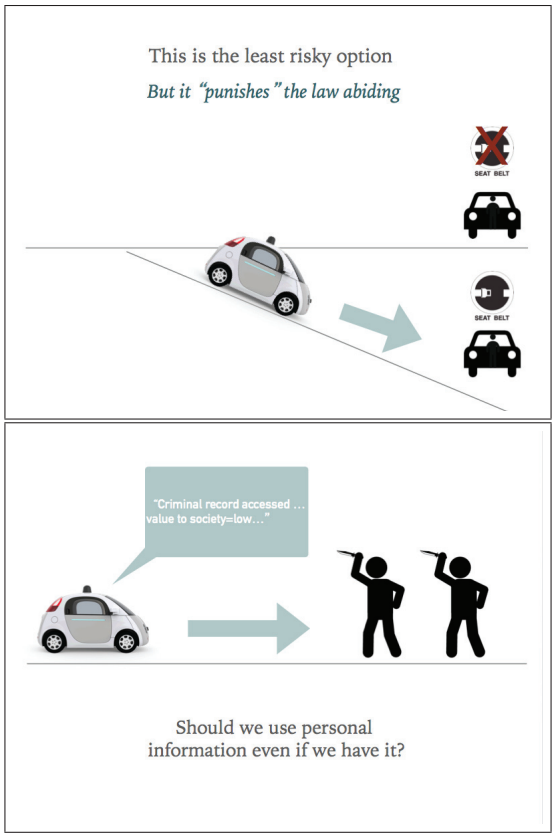

Figure 2: Example slides from lecture

passenger in autonomous car dies vs. five pedestrians die), 18 student pairs simply applied the principle that losing fewer lives is better, so the passenger should die. But eight pairs pointed out other considerations: cars have a "duty to the passenger"; that the passenger trusts the car to "protect his life"; that consumers would not buy cars that did not have these properties.

Discussion and Analysis of Activity

Once the students completed the exercise, we invited them to share their answers with the class. Because students inevitably disagreed about the cases, a lively debate was sparked about targeting algorithms and about Utilitarianism itself. Students disagreed about what choices ought to be made in more controversial cases, as was explained above. Students also questioned how to apply the basic Utilitarian theory in some cases. For example, students were u(where the car should go) nsure of how to apply Utilitarianism in cases that involved probabilities instead of certainties. Some students also supplied additional test cases, which encouraged further debate. For instance, one student described a case in which an autonomous vehicle would have to choose between a person of political importance, such as a president of a country, and five ordinary citizens.

After this discussion, we gave a second lecture that explained what we can learn from examining thought experiments such as the ones given in class. We began by explaining that one purpose of conducting thought experiments in ethics was to test the strength a particular theory such as Utilitarianism. In examining a range of cases, students were able to see that, although the basic version of Utilitarianism seemed to work well in simple cases there were potential problems in applying it to more controversial cases. We also explained that another purpose of examining thought experiments in ethics is to try to isolate which factors are relevant to moral decisions. For instance, the basic version of Utilitarianism assumes that the only factors that are relevant to moral decisions are the amount of benefit or harm they produce. However, certain test cases challenge this assumption. Many students objected to the fact that the basic version of Utilitarianism seemed to require that the vehicle target a person wearing a seatbelt over a person not wearing a seatbelt because this decision would be more likely to minimize harm. It could be argued that the decision, though one that maximizes utility, is unjust because it essentially penalizes an individual for being responsible and for obeying the lawful. If that is the case, then perhaps justice is a potentially morally relevant factor which must be accounted for in the development of a targeting algorithm—in place of, or in addition to, utility.

In the final part of the lecture, we walked students through a series of potential problems with straightforwardly applying a utilitarian approach to developing targeting algorithms for autonomous vehicles—most of which could be explained with reference to the test cases the students had worked through. We then briefly explored some alternatives to basic utilitarianism including modifications of the theory and alternative moral theories. We emphasized that although we had not solved the question of how to program autonomous vehicles, we had come closer to a solution in that we had a better understanding of the ethical issues at play.

Term Project

The course included a four-week term project; the module on ethics was presented to students immediately before commencement of project work. The objective of the term project was for students to apply AI theory that they had learned in a project of their own choosing. Students worked primarily in teams of two, with a few teams of three. For their project papers, students were given this prompt:

> In the Discussion, include one or more paragraphs commenting on the ethical implications of your project.
>
> Some examples of student work include:

- In a project that used neural networks to learn strategies in the video game Doom, students discussed the issue of drone attacks, considering both remote-controlled and autonomous aerial vehicles.
- In a project that optimized the tool path of a manufacturing process for a missile nose cone antenna, students discussed the ethics of weapons design. (This project was an extension of an existing undergraduate research project.)
- In a project that aimed to automatically fact-check statements made in news stories, students noted the importance for the AI to show its reasoning—a concern high-

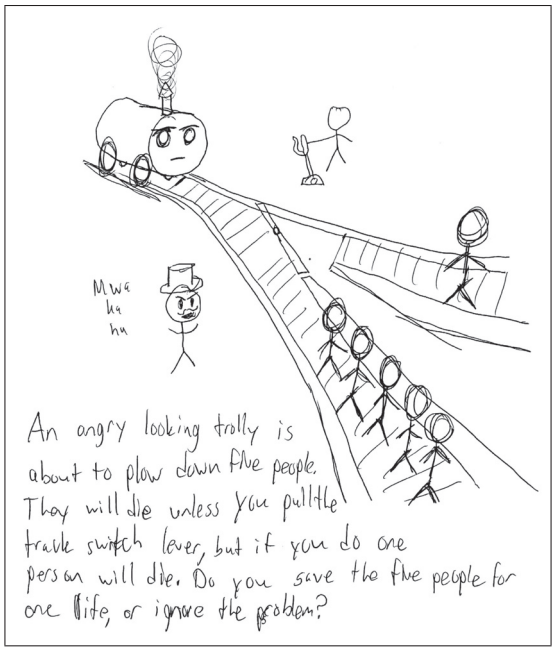

Figure 3: Student explanation of Trolley Problem

lighted by Bostrom and Yudkowsky in their discussion of the ethics of credit reporting algorithms (2014).

Another project made a specific connection between the nature of the AI algorithm used and its implication should it be used in the real world. The project used q-learning to develop a strategy for playing the classic board game Battleship. The students noted that q-learning necessarily explores unknown strategies as part of its learning process. The ethical issue arises if this learning has to occur when the agent is interacting with the real world. The students raised the following ethical dilemma:

> What if the AI's choices lead a search party away from the person they are searching for? Would this be at all acceptable for the AI to do this while it is learning, so that it may be able to save hundreds more once it has learned?

While students did not connect their ethical analyses to Utilitarianism, they did take seriously the charge to reason about the ethical implications of their projects.

Final Exam

In the course final exam, one of the nine problems quizzed the students on the Trolley Problem. Students were given this question:

> Briefly explain the Trolley Problem. You must draw a diagram as part of your answer. The written part of your answer should explain the diagram. You do not need to write a long answer—please explain only the key idea.

62 of the course's 63 students answered correctly. An exemplary answer is shown in Figure 3.

Concluding Discussion and Future Work

The results from our work suggest that it is feasible to introduce students to ethical thinking using a one-week module as part of a semester-long AI course. Students enjoyed the lecture material and found the in-class exercises engaging. With only a minimal prompt, they made reasonable attempts to consider the ethical implications of their project work. They had nearly perfect performance on an example question which asked them to describe the Trolley Problem. In short, we consider these as good results for a relatively small investment of course time.

There are some easy things to make improvements in a future offering. The final project guidance could be strengthened to encourage students to make deeper connections between ethics and AI. We could share the example of the students who questioned whether it would be ethical to allow their AI (q-learning) to explore in the real world, and thereby fail to do good (or do harm) in expectation of improved future performance. We would ask students to make explicit the connections between their AI algorithm and impacts in the world. Students could also be asked to make direct connections between their project and Utilitarian theory.

The alignment between the normative approaches of AI and Utilitarian theory is quite strong. In AI, we typically create value functions which capture the "goodness" of a given world-state. A typical value function is a weighted sum of feature-extractors over the world state. This approach collapses a world-state into a single-dimensional quantity whose value is to be maximized.

As noted earlier, Goldsmith and Burton remarked that Utilitarian theory is "most compatible with decision-theoretic analysis" of AI (2017). A stronger statement would point out that, in a profound sense, normative AI approaches *presuppose and reify* the utilitarian mindset.

For this reason, we have some concern that rather than revealing the limitations of Utilitarianism, it's possible that by naming it, we further reinforced these normative decision-theoretic approaches used in AI.

Several of the test cases we introduced to students were designed to call into question the assumptions of Utilitarianism; e.g., considering passengers who failed to wear seat belts raised issues of justice. By examining these cases, we intended that students gain an understanding of the complexities of moral issues in technology. The question of which strategy to employ—utilitarian or otherwise—is itself a substantive moral question.

To ensure students recognized the tensions raised by Utilitarianism, we would invite students to create equations or code to represent their solutions to the test cases—and highlight the places where this cannot be done.

Acknowledgments

The authors would like to thank the three anonymous reviewers, each of which made helpful comments on the paper draft. We would like to thank our students for engaging with us. This material is based upon work supported by the National Science Foundation under Grant No. SES–1734521.

References

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica,* May 23.

Bonnefon, J.-F.; Shariff, A.; and Rahwan, I. 2016. The social dilemma of autonomous vehicles. *Science* 352(6293):1573–1576.

Bostrom, N., and Yudkowsky, E. 2014. The ethics of artificial intelligence. In Frankish, K., and Ramsey, W. M., eds., *The Cambridge Handbook of Artificial Intelligence*. Cambridge University Press. chapter 15, 316–334.

Burton, E.; Goldsmith, J.; Koenig, S.; Kuipers, B.; Mattei, N.; and Walsh, T. 2017. Ethical considerations in artificial intelligence courses. *arXiv preprint arXiv:1701.07769*.

Burton, E.; Goldsmith, J.; and Mattei, N. 2015. Teaching AI ethics using science fiction. In *AAAI Workshop: AI and Ethics*.

Burton, E.; Goldsmith, J.; and Mattei, N. 2016. Using "The Machine Stops" for teaching ethics in artificial intelligence and computer science. In *AAAI Workshop: AI, Ethics, and Society*.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv preprint arXiv:1703.00056*.

Citron, D. K., and Pasquale, F. A. 2014. The scored society: Due process for automated predictions.

Datta, A.; Tschantz, M. C.; and Datta, A. 2015. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies* 2015(1):92–112.

Furey, H., and Martin, F. 2018. A module on ethical thinking about autonomous vehicles in an AI course. In Neller, T., ed., *EAAI Model Assignments*. http://modelai.gettysburg.edu/.

Garcia, M. 2016. Racist in the machine: The disturbing implications of algorithmic bias. *World Policy Journal* 33(4):111–117.

Goldsmith, J., and Burton, E. 2017. Why teaching ethics to AI practitioners is important. In *AAAI*, 4836–4840.

Nourbakhsh, I. 2017. Ethics and robotics: A teaching guide. https://www.sites.google.com/site/ethicsandrobotics. Accessed September 22, 2017.

Tesla. 2017. Autopilot. https://www.tesla.com/autopilot. Accessed September 22, 2017.

Thomson, J. J. 1985. The trolley problem. *The Yale Law Journal* 94(6):1395–1415.

Van de Poel, I., and Royakkers, L. 2011. Normative ethics. In *Ethics, Technology, and Engineering: An Introduction*. Wiley-Blackwell. 1991–2042.

Wollowski, M.; Selkowitz, R.; Brown, L. E.; Goel, A. K.; Luger, G.; Marshall, J.; Neel, A.; Neller, T. W.; and Norvig, P. 2016. A survey of current practice and teaching of AI. In *AAAI*, 4119–4125.