Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Why Teaching Ethics to AI Practitioners Is Important

Judy Goldsmith, Emanuelle Burton

Abstract

We argue that it is crucial to the future of AI that our students be trained in multiple complementary modes of ethical reasoning, so that they may 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.

Introduction

Consider the decision many of us made last year about whether to sign the open letter petitioning the UN to ban the development of weaponized AI. The authors (one of whom is an AI practitioner who also had to make this decision) use this case study to demonstrate that a knowledge of ethical frameworks is a crucially important tool in an AI student, AI practitioner, and AI theorist's toolbox.

For some AI practitioners, the decision to sign this open letter was a no-brainer, either because those individuals were already committed to non-violence in some form, or because they had thought at length about the dangers of weaponized AI. Some no doubt signed because the leaders of the community did so, and they wanted to be seen as one of the "cool kids." Many others, however, chose not to sign the letter, and even strongly opposed it, because they believe that weaponized AI is inevitable, or desirable; because they believe that any AI can be used as a weapon;[1] because they were reluctant to associate their name with a petition when they could not predict the ramifications, or for many other reasons arising from their understanding of research, politics, or their personal moral imperatives.

It was not entirely clear where a greedy agent would land on this question. The letter was linked to an organization that was offering grants for AI, more particularly for ethics-and-AI projects. On the other hand, one might reasonably conclude that signing such a letter could have a negative effect on possible support through military-funded research. In some countries, there is significant military funding for AI, albeit often for work couched in terms of defense and security, rather than offense and weapons. Some research labs are able to support students because they accept military funding; as such, concern about funding cannot be dismissed as purely selfish (though they still reflect a preference for their own community of AI researchers.)

In fact, many people made their decision based on moral concerns, and yet arrived at different answers. It wasn't just that people were coming to different conclusions; in many cases, they were beginning with different ideas about how to make an ethical decision. In this paper, we show how knowledge of different ethical frameworks can illuminate the different approaches to decision-making that different AI practitioners took, and re-examine the question of whether or not to sign from within each of these frameworks.

We demonstrate that most AI practitioners operate within the ethical framework called utilitarianism, which has been the dominant mode of ethical thought in the west for the past 150 years, and which is the ethical theory that is by far the most compatible with decision-theoretic analysis (Burton et al.). After describing utilitarian theory, we briefly introduce deontology and virtue ethics, the other two major modes of ethical analysis, and show how these two modes can offer new perspectives on the decision of whether to add one's name to such a letter. We readily acknowledge that there is much more to AI than war bots, and more to practitioners' decisions about public declarations than this initial summary indicates. We could apply similar framing and analysis to the use of AI in medicine, management, computer games, or any other area. However, this case exemplifies one of the two broad types of decisions that call for ethical analysis: personal decisions by AI practitioners and programmers, and decisions made by AI systems.

What appears to be a simple binary decision — sign the letter, or not? — is only the final stage in assessing one's (probably non-binary) views on several complicated questions. Should robots be used to kill people? Under what conditions could such robots be developed responsibly, and are those conditions in place? But also: what other valuable or positive purposes could this same technology serve? Could our work in other areas of AI continue without the financial support of the military? For many AI practitioners, the answers to these questions are not black and white.

[1] In the words of Ani DiFranco, "'Cause every tool's a weapon — if you hold it right." (From the song My I.Q.)

A Utilitarian Analysis of the Decision to Sign

When making a decision, computer scientists usually analyze the question in terms of utility. The first question an AI practitioner might ask, in defining the utility of signing and of not signing, is "Utility to whom?" This agent can consider the cost to herself in terms of potential future military funding, and to her institution and her students. She can weigh that against the effect on her reputation as an ethical researcher. Or she can consider the possible impact on enemies of her country if she were to choose to develop weaponized AI, and weigh that against the impact of having human soldiers attempt the same acts. She might, further, consider the impact of her country — or all countries — having the technology she might develop.

In order to compute the expected values of signing or not signing, this practitioner must decide the values of individual lives that could be ended by the technology, or lack thereof. She must decide whose utility matters, and the relative weights of each person's needs and desires. She must decide how she will handle expected future rewards and payoffs. Each of these modeling decisions has enormous impact on the optimal policy for that model.

This type of moral reasoning has a solid foundation in ethical theory, specifically in a theory called utilitarianism. As ethical theories go, utilitarianism is very recent, dating back only to the late 18th century (though some elements of it appeared much earlier. (Driver 2014).) Its basic principle is commonly formulated as "the greatest good for the greatest possible number." Utilitarianism holds that one's primary moral obligation is to work for the greatest possible happiness (defined by John Stuart Mill in his influential book *Utilitarianism* both as pleasure and as the absence of pain) for the greatest number of people (Mill 2002). This insistence on public (or even universal) good is what distinguishes utilitarianism from a simple cost-benefit analysis, because it demands that the agent discount her own preferences for herself, or for certain favored groups.

Note that the language "optimal policies" in the example above corresponds to that of decision-theoretic planning; we perceive a strong correspondence between a utilitarian analysis of a decision on the one hand, and modeling the decision and consequent choices as a Markov decision process on the other. In either case, the agent is attempting to optimize total expected utility over time. Deciding on the scope of utility, and assigning values to different kinds of lives, are the "moral" decisions available. Everything else is descriptive, although its validity as description depends on first granting the values the agent has assigned.

In contrast to the two other major schools of ethical theory, utilitarianism is concerned only with outcomes, rather than with methods or intentions. This means that any possible law or rule could be set aside, in a given situation, if an agent determines that adhering to that law conflicts significantly with the greater good. According to utilitarianism, an agent is sometimes *required* to do harm in order to choose the best possible course; as such, the agent is not morally accountable for harm done under such circumstances, because it was the "right" choice. "Ticking time bomb" scenarios, such as those portrayed in the TV series *24* (in which counterterrorism agent Jack Bauer routinely tortures suspects for information), highlight both the prevalence of utilitarian thought and the appeal of this line of reasoning (Nissel 2010).

But decision-theoretic analysis is not the *same* as utilitarian analysis, and applying the core utilitarian principle of "the greatest good for the greatest possible number" places some limits on my decision-theoretic analysis that are both challenging and useful. This principle requires that the moral agent consider the well-being of everybody who is even potentially affected, and act in the way that produces the greatest possible benefit across that group. It is not acceptable, for instance, for her to decide that the needs of AI researchers (or denizens of her country, or members of her own faith) outweigh the needs of other people; and if she finds herself reaching this conclusion, she needs to submit it to careful scrutiny. What is less obvious is how different kinds of utility (running the gamut from basic physical safety to improved professional opportunities) should weigh against each other, and — even more importantly — whether the well-being of "the enemy" should figure into her calculations, and how much their well-being should matter compared to her own side's citizens and soldiers.

Utilitarian thought — which we have only briefly summarized here — does not provide straightforward answers to these questions, but it offers analytical tools to help one address them thoroughly and responsibly in a range of situations. It can enable the AI practitioner to reach a more ethically comprehensive position, allowing her to deploy familiar modes of reasoning while challenging her to look beyond her own utility and personal concerns.

Despite the value of utilitarian ethics, we believe that AI practitioners should also be familiar with the other two major schools of ethical theory, deontology and virtue ethics. These two approaches are far less compatible with decision-theoretic analysis and other familiar analytic strategies, which can make them challenging to understand and to apply. We argue that it is worthwhile, even essential, for AI practitioners to confront this challenge, and apply these theories in order to achieve the clearest possible understanding of a given situation, and of their own reasoning and decision-making in response to it.

Deontology: Ethics by Rules

Though utilitarianism is the ethical theory most compatible with contemporary culture, and thus typically feels the most "useful", almost everyone in the world today also has some experience with deontology, or law-based ethics. Deontological ethics (from the Greek "deon," which means "duty" or "obligation"), conceives of ethics in terms of laws, or rules: my actions are ethical insofar as they conform to (or do not violate) the law (Alexander and Moore 2015). It is worth noting that there can be several layers of law; for example, the law "do not kill" is more fundamental than the law to drive on the correct side of the road in your country, but this latter law is still binding (except in those rare conditions when it is necessary to violate the lower law to preserve the higher one), because it creates the conditions for you to follow the moral law. Though is not always easy to know what

the law requires, one is always required to follow it. This requirement makes deontology far less flexible than utilitarianism, but can also furnish one with the conviction to take difficult or unpopular stances.

There are different ways of understanding or defining the law. The three Abrahamic religions (Judaism, Christianity, and Islam) are all Divine Command traditions, in which the law is understood to be given by God; it is a person's duty to follow that law, although it is recognized by most denominations of each of these religions that human beings have to do a lot of work to interpret the laws, and to ascertain the best way to apply them in complex situations. In Immanuel Kant's reformulation, however, the moral law is something each individual must discover herself, not by trusting in authority but through the ongoing application of reason. According to Kant, the true law is universal: the further a given principle can be generalized, the closer it is to the true law (Rohlf 2016). (Thus, one can discover through reason that "do not kill" is more fundamental than "drive on the correct side of the road.") In all forms of deontology, the ability to follow the law relies on the agent's ability to correctly assess what part(s) of the law are most fundamental.

The language of rules or laws can make deontology seem analogous, at first glance, to the application of axiomatic systems. In some ways, the comparison is helpful: just as members of religious traditions have to analyze situations that fit uncomfortably within existing laws (sometimes because the situations are the product of modern developments that postdate the laws (see extensive discussion of this in Johnson (2009)), sometimes because the situation exists at the intersection of several laws, which appear to dictate different solutions), so too do programmers spend considerable time, energy, and creativity on exception handling. But abiding by deontology is more like living within an axiomatic system than building one, because deontological reasoning and analysis do not allow you to change the laws. Whether the laws are given by authority or ascertained by reason, they are understood to exist independently of the goals or desires of the individual.

Although deontology presumes that a given agent abide consistently by the same set of laws or axioms (rather than picking and choosing according to the situation), one can begin with any somewhat-general axiom and apply deontological reasoning. Consider Aaron Swartz' project to make copyrighted articles freely available (Mechanic 2013; Swartz 2008). One could argue that, since Swartz was living in the United States and benefiting from the order created by its laws, he was bound to uphold those laws; and because downloading copyrighted material with intent to share it freely is illegal, Swartz's actions were therefore unethical. Alternately, one could begin with the axiom that scholarship should be publicly available (or, similarly, that scholars should be able to distribute their own work freely), and that Swartz acted rightly; indeed, if one considers the free availability of scholarship to be a fundamental law, then one is obligated to uphold that law even if one will face criminal charges. Deontological analysis helps explain why Swartz might have felt his actions were ethically necessary; whether or not one believes they were ethical, full stop, will depend on whether or not one shares his axioms, and believes it is right to apply those same axioms in all situations concerning intellectual property, or violations of civil law.

Similarly, there are different possible starting places for a deontological analysis of the decision to sign (or not to sign) the open letter. An AI practitioner who begins with the law "do not kill" will probably sign the letter, because building war bots requires working against that principle. But he also might also decide not to sign on the basis of Just War Theory. This theory invokes cost-benefit logic to argue that war, while awful is sometimes better than the alternative. Just War Theory demands that specific criteria be met in order for a war to be deemed just; for example, acts of war must be proportional, and must be directed only at active enemy combatants, rather than civilians or injured soldiers. If this practitioner were to conclude that military AI would help his country's military effect more targeted attacks with less collateral damage (and if he trusted military leaders to use it in that way), then he might conclude that developing this technology would be the best way to honor this principle. Another practitioner might believe that her primary duty is to fulfill her professional duties because she has made the commitment to do so, both by becoming an expert in the field and by accepting a job. She may conclude that she should not sign the letter, because so much of AI research relies on military funding. Alternately, she might decide that AI research needs to break its ties to the military, even though this break will entail a significant loss in funding and prestige; if she reached this conclusion, she would be morally obligated to sign onto the letter, even though the consequences would be severe.

These are only two of many possible principles or axioms that could be used to guide one's reasoning. As we have shown, deontological reasoning does not always lead different agents to the same conclusions, even when the foundational principle is consistent, but will also depend on how the agent describes the situation (s)he is responding to (e.g., will this technology be used to pursue a more just war?) For the practitioner struggling to ascertain which principle(s) should take priority over the others, deontology's preference for generalizable rules can be a helpful way to determine which principles are most fundamental.

Virtue Ethics

Virtue ethics, the third major approach to ethics, is focused on individual character, and how one develops good qualities"or virtues" (such as honesty or courage) and the ability to apply them. Virtue ethics has ancient roots in both Greek and Chinese thought, though contemporary terminology is drawn primarily from Aristotle, whose *Nichomachean Ethics* is still considered foundational (Hursthouse 2013) (Aristotle 1999). Virtue ethics is a "big picture" system, and individual actions and problems are evaluated in terms of how they fit into the arc of a person's life. According to virtue ethics, a person's character is a product of habits, which are strongly influenced by their social context: a person is much more likely to do something that feels ordinary, whether because it matches their self-understanding or just seems like a "normal" thing to do. This means that a person's

past actions are a useful predictor of how they will choose to act in the present. It also means that a person's present-day choices can and should be understood, in part, as choices about who she will become in the future, because they affect the scope of actions that feel familiar or "like me".

Virtue ethics offers a very different framework for a practitioner who is deciding whether or not to sign the letter. Because virtue ethics is concerned with character, it focuses on long-term patterns of action, rather than on single acts; the analysis of any single decision will need to be framed by questions like: "who do I want to be?" and "what do I hope to accomplish?" The agent's decision might be entirely separate from his beliefs about whether an open letter to the UN will impact the future of mechanized warfare, but will instead reflect what he believes will be his decision's impact on his own sphere. His question is not whether robots should kill, but whether he should be the sort of person who publicly protests against weaponized AI, and how the decision to sign will influence the person (and the professional) he will become. He will probably weigh the question of whether signing the letter will endear him to his supervisor and colleagues or alienate them; he might also consider how his decision will affect the possibility of future avenues of research, though he might conclude that his decision about the letter is more important. He might consider how his decision will raise his profile, within his local community, as an advocate for certain kinds of causes. He might ask himself how he would explain his decision to his spouse, or his friends, and how it would change their understanding of him. His decision may also depend on whether he opts to define himself as an AI professional, as a US citizen, as a father, or as a human being (whatever any of those terms means to him.) All of these questions reflect the practitioner's interest in who he will become, both as an individual and as a member of his community. Many people already ask themselves these sorts of questions when facing an important decision. Virtue ethics offers a model for organizing them around considered goals, and for evaluating how each possible choice fits in with those goals.

By focusing on the social context of ethical action, virtue ethics also affords a framework for an agent to reflect upon the ways in which AI, and the actions of AI practitioners, creates a broader social context for moral action. Virtue ethics prompts us to consider the character of the decision-makers who program war bots, or the politicians or military leaders who decide to deploy them. According to virtue ethics, the act of making these decisions, and of engaging in their implementation, has an impact on those individuals: over time, their character shifts to accommodate the actions they are undertaking, as their sense of what is normal or appropriate migrates. By building robots that kill (even if only for particular specialized purposes) they adapt their sense of what is normal to make room for what they are doing, and thus they become the sort of people who are more likely to think robots that kill are the appropriate solution to a given problem. Virtue ethics therefore pushes us to consider the sort of decisions that are inherent in the jobs AI practitioners have, and helps us think about the social and technological contexts that force such choices. This can help us see connections between technological development and moral choices, and to choose whether or not to act in ways that will protect our own and others' characters.

The virtue ethics perspective shows us that developing war bot technology changes the social conditions for moral decision making. By engaging with the virtue ethics framework, AI practitioners can interrogate how AI shapes society and thus creates the conditions for certain moral crises.

Conclusions

We have used the question of choosing whether to sign an open letter to show that applying ethical frameworks to the analysis of a decision greatly enriches our decision processes, and gives us tools for understanding and evaluating decisions. As the above discussions make clear, ethical theory introduces new critical tools for analysis, but can help identify the ways in which one is already thinking and reasoning about ethical questions. This same way of reframing and discussing decisions can apply to many kinds of decisions we make as AI practitioners, and to decisions our software may make on our behalf (Burton, Goldsmith, and Mattei 2015).

Because AI work has enormous effects on society — including the ways in which individuals interact, on our economies, on the practice of medicine and the uses of leisure, to name a few — we believe that all AI practitioners, and those that reason about AI technology, should be able to frame discussion in terms of ethics. Further, we believe that popular culture promotes a very limited understanding of how to frame and analyze ethical decision making. Thus, we advocate that ethics be taught in AI classes, and that we develop materials and courses for teaching ethical frameworks and reasoning to people working in AI (Burton, Goldsmith, and Mattei 2016). (See also (Burton, Goldsmith, and Mattei 2015; 2016) for a discussion on using science fiction to each AI ethics.)

Acknowledgments

This material is based upon work 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(s) and do not necessarily reflect the views of the National Science Foundation. We thank Nicholas Mattei (our collaborator in the long-term project that informs this paper) and John Fike for helpful discussions.

References

Alexander, L., and Moore, M. 2015. Deontological ethics. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Spring 2015 edition.

Aristotle. 1999. *Nichomachean Ethics*. Hackett. trans. Terence Irwin.

Burton, E.; Goldsmith, J.; Koenig, S.; Kuipers, B.; Mattei, N.; and Walsh, T. Ethical considerations in artificial intelligence courses. *Invited for AI Magazine*.

Burton, E.; Goldsmith, J.; and Mattei, N. 2015. Teaching AI ethics using science fiction. In *1st International Workshop on AI, Ethics and Society, Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Burton, E.; Goldsmith, J.; and Mattei, N. 2016. Using "The Machine Stops" for teaching ethics in artificial intelligence and computer science. In *AI & Ethics: Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

Driver, J. 2014. The history of utilitarianism. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Winter 2014 edition.

Hursthouse, R. 2013. Virtue ethics. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Fall 2013 edition.

Johnson, D. G. 2009. *Computer Ethics*. Prentice Hall Press.

Mechanic, M. 2013. Steal this research paper! (You already paid for it.). *Mother Jones*.

Mill, J. S. 2002. *Utilitarianism*. Hackett Publishing Co.

Nissel, M. 2010. The ever-ticking bomb: Examining 24's promotion of torture against the background of 9/11. *aspeers* 3:37–51.

Rohlf, M. 2016. Immanuel kant. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Spring 2016 edition.

Swartz, A. 2008. Guerilla open access manifesto. Online-Ressource, http://archive.org/stream/GuerillaOpenAccessManifesto/Goamjuly2008_djvu.txt.