Assessing Ethical Thinking about AI

Judy Goldsmith,[1] Emanuelle Burton,[2] David M. Dueber,[1]
Beth Goldstein,[1] Shannon Sampson,[1] Michael D. Toland[1]
[1]University of Kentucky
[2]University of Illinois at Chicago

Abstract

As is evidenced by the associated AI, Ethics and Society conference, we now take as given the need for ethics education in the AI and general CS curricula. The anticipated surge in AI ethics education will force the field to reckon with delineating and then evaluating learner outcomes to determine what is working and improve what is not. We argue for a more descriptive than normative focus of this ethics education, and propose the development of assessments that can measure descriptive ethical thinking about AI. Such an assessment tool for measuring ethical reasoning capacity in CS contexts must be designed to produce reliable scores for which there is established validity evidence concerning their interpretation and use.

Advances in computer technology continue to change the conditions of human life, and thus of ethics, for everyone. Yet, computer scientists bear a particular responsibility in this changing ethical landscape: what we build can create new possibilities for kindness and justice. Or, what we build can hinder them. Given that AI has had such an impact on the ethical terrain we are still trying to figure out how to teach AI ethics effectively. Concurrently, if we are to evaluate the quality of curricula, we need a clear definition of effectiveness and then assessments that can produce valid and reliable scores. Thus, we propose that the CS community should develop an assessment tool to measure students' ethical reasoning and descriptive insight skills applied to CS and, specifically, AI related issues. The new assessment would be designed as a tool that could be used to evaluate instruction geared towards teaching CS ethics.

We argue that a focus on effectiveness of AI ethics instruction is important and timely; today's undergraduate CS education is preparing tomorrow's technical professionals. AI professionals inevitably encounter ethical challenges of some kind in the workplace. The ethical burden on responsible professionals in AI is especially pronounced. In striving to design technology that supports ethical living and ethical social structures, these professionals are faced with profound and particular dilemmas that arise from the ubiquity of computation in all aspects of modern work and life. They must consider the individual, social, economic, political, and environmental costs of the design, manufacture, and use of the technology, while being able to make mindful judgments on their own. The stakes are high, as AI professionals design the machines, controllers and software for so many aspects of our lives, including manufacturing, healthcare, entertainment, transportation, education, policing, and war.

Our students are grappling with technologies such as mass surveillance, big data methods, social network analysis, neurologically controlled prostheses, robot caregivers, and self-driving cars, just to name a few. Because these technologies are new and constantly emerging, and their impact on the world is not fully understood, it is neither possible nor useful for a CS ethics curriculum to try to offer students a comprehensive account of the tech ethics landscape. Rather, an effective curriculum in AI ethics must train its students to describe for themselves, carefully and clearly, both emergent technologies and the aspects of life that they will affect. Indeed, some of the most pressing ethical issues AI students will confront will not initially appear to be ethical problems. As new technologies and applications are developed, it will fall to our students to identify, describe and address potential ethical dilemmas that their colleagues may not recognize as problems.

Particularly in a field so future-oriented as AI, any normative judgments about how to design and implement these technologies are likely to go awry if they are not founded in clear-eyed and deliberate description. Accordingly, CS ethics education (and AI ethics education in particular) should not only develop the skills in students to be able to analyze the ethical strengths and weaknesses of existing computer technology but should empower them to imagine intended uses and possible ramifications of technology before it is released into the world. Furthermore, they must have the capacity to reason through how they might resolve a problem (Burton, Goldsmith, and Mattei 2018; Goldsmith and Burton 2017; Quinn 2006).

In order to gauge the extent to which CS programs are preparing students to meet these challenges, we need an assessment tool that can measure whether these components are being transmitted to students effectively. Gathering this information is more challenging than testing for students' knowledge of programming languages or software design. How can we know a given person's capacities for ethical reasoning? How can CS ethics instructors measure their students' sense of commitment to ethical decision-making? Developing an assessment tool that is easily usable and informative will require extended development time and interdisciplinary collaboration between AI experts who know the field of AI ethics and its student population, and testing and measurement experts who know how to develop and evaluate the quality of such assessment tools. Consequently, individual instructors are unlikely to develop such a tool unless they have training in test development and psychometrics.

Burton and Goldsmith each teach a variant of a course, Science Fiction and Computer Ethics (Goldsmith and Burton 2017; Burton, Goldsmith, and Mattei 2018). In that course, students use ethical theory both to understand the ethical quandaries portrayed and implied in the stories and to address them. Students are challenged to consider the state of technology described in the fiction, and the interactions between technology and society. They explore how these fictional portrayals elucidate their own present (and immediate future). The students work descriptively with major ethical theories and have the opportunity to develop a rich critical vocabulary for recognizing ethically fraught situations. By the end of the course, students should be able to identify potential ethical risks in a given technology or model, or in a company's and the public's use of this technology or model. They should be able to recognize potential ethical pitfalls in a given project, to articulate the costs and benefits of each potential solution, and be able to identify and critique incomplete or specious ethical justifications.

This set of correlated skills is not easy to teach, let alone to assess, and existing literature on AI ethics education focuses almost exclusively on teaching, rather than on assessment. Additionally, of the few assessments available for measuring ethical development, none sufficiently operationalize ethical reasoning skills as we suggest AI needs. Such an assessment should be driven by a thoughtfully constructed conceptual definition of "ethical reasoning skill," formed with input from faculty, students, and professionals in AI and computer ethics fields. Development of the assessment should be a collaborative effort involving experts in computer science and ethics working with those experienced with developing assessments and evaluating the reliability and validity of assessment scores using modern measurement techniques.

The Development of Ethical Reasoning

The domain of ethical reasoning capacity can be seen as consisting of two facets or domains: ethical description and felt ethical responsibility. The relationship between these two domains is complex. It is of course possible to have a highly developed sense of responsibility but weak ethical description ability.I t is also possible to be highly capable of ethical description but feel very little ethical responsibility to help others or improve the world.

We suggest that both of these capacities should be assessed and evaluated. Our goal in suggesting that these capacities be assessed in tandem is not to simplify the complexity of human experience, but rather to identify the cultivation and integration of these two capacities as a crucial goal of AI ethics education. There is also a good basis, descriptively speaking, for understanding them to be deeply related. While the ability to describe ethical issues is not coextensive with feelings of ethical responsibility, it has been argued as far back as Plato (Plato 2006) that true, genuine understanding of what is good necessarily involves the ability, or at least the desire, to be good. While we do not purport to offer definitive answers to this age-old philosophical debate, we concur that ethical understanding is not purely informational. Because ethics is concerned with what ought to be, as well as what is, a thorough ethical reasoning assessment needs to take account of the way in which the student's ability to perform ethical description is connected to their sense of ethical responsibility.

Existing Ethics Assessments

There do exist some assessment tools for ethics education. However, they are framed by assumptions about the nature of ethics and ethical decision-making that are not aligned to the outcomes we propose are important to CS ethics.

In this section, we present the most commonly used assessments, discuss what they are designed to measure, and explain why the existing assessments fall short of capturing the critical outcomes of ethics education. We then describe the possibilities for an assessment better targeted to the ethical quandaries a CS professional is likely to encounter, recognizing that the ethical dilemmas of the future are not easily captured in present-focused scenarios. The unique need in CS ethics to develop present ethical thinking in preparation for an unknown technological future makes field-specific ethics assessment particularly important.

The Defining Issues Test (DIT) and DIT-2 have traditionally been used to measure short-term ethical development in NSF-funded projects, but these measures have been criticized as being too general in focus to be used in specific settings (Titus et al. 2011; Woodward 2007). The DIT-2 is based specifically on Kohlberg's theory of ethical development (Rest et al. 1999a; Thoma and Dong 2014; Zhu et al. 2014), which is framed in deontology, one ethical paradigm among others. Any ethical metric that is wedded to a particular system will fail to acknowledge the legitimacy and necessity of multiple approaches. However, Kohlberg's theory of ethical development has been specifically critiqued for privileging one single model of ethics over other legitimate models (Flanagan 2009). Furthermore, (Zhu et al. 2014) acknowledge that Kohlbergian-based assessments may not account for the different kinds of ethical reasoning required by engineers in their design processes. A number of assessments based on the DIT-2 have been developed specific to a particular domain: two examples within the field of engineering are the *Engineering Ethical Reasoning Instrument (EERI)* (Zhu et al. 2014) and the *Engineering and Science Issues Test (ESIT)* (Borenstein et al. 2010). In keeping with the DIT's Kohlbergian foundations, these assessments are designed to measure moral judgment (Bebeau 2002) in deontological terms, and as such are not equipped to recognize or assess other modes of ethical engagement. This approach is, at best, incomplete; some go further, and argue that the primary benefit of ethics education is that it emphasizes less moral judgment and more ethical sensitivity, the ability to recognize relevant ethical issues emerging from a situation (Drake et al. 2005). Furthermore, as with the DIT-2, the EERI and ESIT items consist of ranking the relevance of specified ethical issues within the presented scenario. Since the items only assess recognition and offer no open or constructed response element, they cannot fully measure competence in ethical reasoning, which we have argued is essential. The Test for Ethical Sensitivity in Science and Engineering (TESSE) is designed to measure ethical sensitivity (Borenstein et al. 2008), but development of this test was never completed, and it is unpublished. Furthermore, the example scenario provided by Borenstein et al. (Borenstein et al. 2008) concerns professional ethics rather than ethical considerations related to impact on society. For these reasons, the TESSE is not appropriate for our purposes, nor for the purposes of the emerging needs of ethics education. Thus, while assessments for measuring ethical development exist, they are typically narrowly framed by one ethical theory (Zhu et al. 2014). Or, existing assessments focus on only one of many aspects of ethical conduct, arguably a different goal than that of most CS ethics' education courses (Borenstein et al. 2008) and different than our proposed objective to assess ethical reasoning capacity. Indeed, developers of existing assessments call for revision, refinement, and further validation (Rest et al. 1999b; Rest and Narvaez 1998; Rest et al. 1999a; Borenstein et al. 2008).

Why a New Assessment Is Useful

Beyond educational uses, we argue that the development of a more AI-focused ethical-thinking assessment tool has broad implications as a tool to assess ethical reasoning capacity for AI students entering the workforce. The putative assessment can provide a measure of graduate preparedness to navigate our technological future and attend to societal benefits of emerging technologies with respect to ethical implications. In his Ninety-five Theses for Reforming Program Evaluation, (Cronbach and others 1983) Cronbach writes "Whatever the evaluator decides to measure tends to become a primary goal of program operator", but only if the evaluator is in a position of power to influence resources for the operators. Because our blue sky proposal is to design an assessment with end uses including student, curricular, and program evaluation, we recognize the assessment has implications for AI ethics instruction, and all of CS ethics instruction. In fact, the influence of assessment on instruction can motivate us to explore a new way of assessing CS ethics, as existing assessment tools are not well aligned to a nuanced and substantive definition of ethical reasoning capacity.

Conclusion

The push toward teaching AI ethics comes out of a profound understanding that ethical design and analysis of AI is necessary for global social welfare (see, for instance, the ongoing controversies around political manipulation, the weakening trust in online banking, or any of the many hot-button AI-related topics in our news cycles) and perhaps survival (see the near-constant references to the Terminator movies in the popular press). If ethical design and analysis of computer technology is indeed necessary for global social welfare, we posit it is also important to know if students are graduating with an ethical description and reasoning capacity that prepares them to navigate an ever-changing technological future. A carefully designed assessment will allow academic programs and faculty to explore student progress toward this outcome. Further, such an assessment will allow researchers to study the efficacy of different approaches to teaching computer ethics.

Although most of us at AAAI do not have the educational and psychometric backgrounds needed to design effective ethical-thinking assessment tools, we do have the expertise in AI, and a growing awareness of the necessity of ethical thinking about AI. This paper is, therefore, a plea for the teaching of open-ended, description-oriented AI ethics, and a plea for cooperation in the development of assessment tools to measure the effectiveness of our teaching of ethical teaching of AI.

References

Bebeau, M. J. 2002. The defining issues test and the four component model: Contributions to professional education. *Journal of moral education* 31(3):271–295.

Borenstein, J.; Drake, M.; Kirkman, R.; and Swann, J. 2008. The test of ethical sensitivity in science and engineering (TESSE): a discipline-specific assessment tool for awareness of ethical issues. In *Annual ASEE Conference, American Society for Engineering Education, Pittsburgh, PA*.

Borenstein, J.; Drake, M. J.; Kirkman, R.; and Swann, J. L. 2010. The engineering and science issues test (esit): A discipline-specific approach to assessing moral judgment. *Science and Engineering Ethics* 16(2):387–407.

Burton, E.; Goldsmith, J.; and Mattei, N. 2018. How to teach computer ethics through science fiction. *Communications of the ACM* 61(8):54–64.

Cronbach, L. J., et al. 1983. Ninety-five theses for reforming program evaluation. In *Evaluation Models*. Springer. 405–412.

Drake, M. J.; Griffin, P. M.; Kirkman, R.; and Swann, J. L. 2005. Engineering ethical curricula: Assessment and comparison of two approaches. *Journal of Engineering Education* 94(2):223.

Flanagan, O. J. 2009. *Varieties of moral personality: Ethics and psychological realism.* Harvard University Press.

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

Plato, trans. Beresford, A. 2006. *Protagoras*. Penguin Classics.

Quinn, M. J. 2006. On teaching computer ethics within a computer science department. *Science and Engineering Ethics* 12(2):335–343.

Rest, J., and Narvaez, D. 1998. Guide for DIT-2. *Center for* *the Study of Ethical Development, University of Minnesota. Minneapolis, MN*.

Rest, J. R.; Narvaez, D.; Thoma, S. J.; and Bebeau, M. J. 1999a. DIT2: Devising and testing a revised instrument of moral judgment. *Journal of Educational Psychology* 91(4):644.

Rest, J.; Narvaez, D.; Mitchell, C.; and Thoma, S. 1999b. Exploring moral judgment: A technical manual for the Defining Issues Test. *Manuscript available from Center, University of Minnesota*.

Thoma, S. J., and Dong, Y. 2014. The defining issues test of moral judgment development. *Behavioral Development Bulletin* 19(3):55.

Titus, C.; Zoltowski, C. B.; Huyck, M.; and Oakes, W. C. 2011. AC 2011–1833: The creation of tools for assessing ethical awareness in diverse multi-disciplinary programs. In *Proceedings of the 2011 ASEE Annual Conference*.

Woodward, B. 2007. Growth and training impact in IT: A measure of ethical reasoning. *Issues in Information Systems* 8(2):220–224.

Zhu, Q.; Zoltowski, C.; Feister, M.; Buzzanell, P.; and Oakes, W. 2014. The development of an instrument for assessing individual ethical decision-making in project-based design teams: Integrating quantitative and qualitative methods. *Age* 24:1.