Integrating Ethics within Machine Learning Courses

JEFFREY SALTZ, Syracuse University
MICHAEL SKIRPAN, Carnegie Mellon University
CASEY FIESLER, University of Colorado Boulder
MICHA GORELICK, Probable Models
TOM YEH, University of Colorado Boulder
ROBERT HECKMAN and NEIL DEWAR, Syracuse University
NATHAN BEARD, University of Colorado Boulder

This article establishes and addresses opportunities for ethics integration into Machine Learning (ML) courses. Following a survey of the history of computing ethics and the current need for ethical consideration within ML, we consider the current state of ML ethics education via an exploratory analysis of course syllabi in computing programs. The results reveal that though ethics is part of the overall educational landscape in these programs, it is not frequently a part of core technical ML courses. To help address this gap, we offer a preliminary framework, developed via a systematic literature review, of relevant ethics questions that should be addressed within an ML project. A pilot study with 85 students confirms that this framework helped them identify and articulate key ethical considerations within their ML projects. Building from this work, we also provide three example ML course modules that bring ethical thinking directly into learning core ML content. Collectively, this research demonstrates: (1) the need for ethics to be taught as integrated within ML coursework, (2) a structured set of questions useful for identifying and addressing potential issues within an ML project, and (3) novel course models that provide examples for how to practically teach ML ethics without sacrificing core course content. An additional by-product of this research is the collection and integration of recent publications in the emerging field of ML ethics education.

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability;

Additional Key Words and Phrases: Machine learning, ethics, education

ACM Reference format:
Jeffrey Saltz, Michael Skirpan, Casey Fiesler, Micha Gorelick, Tom Yeh, Robert Heckman, Neil Dewar, and Nathan Beard. 2019. Integrating Ethics within Machine Learning Courses. *ACM Trans. Comput. Educ.* 19, 4, Article 32 (July 2019), 26 pages.
https://doi.org/10.1145/3341164

Authors' addresses: J. Saltz, Syracuse University, Hinds Hall, Syracuse University, Syracuse NY, 13244; email: jsaltz@syr.edu; M. Skirpan, Department of Philosophy, Carnegie Mellon University, Baker Hall 161, 5000 Forbes Avenue Pittsburgh, PA 15213; email: mskirpan@andrew.cmu.edu; C. Fiesler and N. Beard, University of Colorado Boulder, UCB 315, Boulder, CO 80309; emails: {casey.fiesler, nathan.beard}@colorado.edu; M. Gorelick, Probable Models, 502 Berlin Road, Pittsburgh, PA, 15221; email: micha@probablemodels.com; T. Yeh, University of Colorado Boulder, UCB 315, Boulder, CO 80309; email: tom.yeh@colorado.edu; R. Heckman and N. Dewar, Syracuse University, Hinds Hall, Syracuse University, Syracuse NY, 13244; emails: {rheckman, ndewar}@syr.edu.

1946-6226/2019/07-ART32 $15.00
https://doi.org/10.1145/3341164

1 INTRODUCTION

The combination of accelerated access to large data sets, improved algorithmic approaches, and advancements in computational hardware and storage methods has created a massive boom in the broad field of Machine Learning (ML). Over a short interval, machine learning techniques have outperformed prior efforts in tasks ranging from image recognition [70] to natural language processing [17] to agent-based systems [94]. However, like many significant scientific advancements, the use of ML techniques has raised a number of significant ethical challenges. For example, ML systems have shown the capability of inheriting racial [1, 19] and gender [3, 27] biases. Further, ML systems have been used to predict, and thus disclose, private attributes of users [48] or even target their beliefs and psychological traits [43, 59].

What brings great power to these new methodologies also yields great responsibility toward the users whose clicks, movements, and social lives feed these systems. As ML systems are deployed deeper into our human systems, such as policing, credit evaluation, news feeds, medical diagnostics, and job applications, we witness the cutting edge of computing diving deeply into people's most personal matters. In effect, ML is establishing a seamless continuum between engineered systems and users' personal actions, beliefs, and well-being. Thus, those who use ML to build applications, solve problems, or conduct research need to be aware that their choices may have profound impacts on others. The ML model they trained or chose to use may give someone a high paying job, grant someone a low interest loan, deny someone parole, or cause a pedestrian to be hit by a car. Thus, to act responsibly, ML engineers must adopt perspectives and competencies that go beyond complexity analysis and usability, and into histories, social sciences, and morality.

Much as Langdon Winner argued many years ago in his seminal essay, "Do Artifacts Have Politics?," technology has social consequences [99], and hence, it is essential that engineers are able to interrogate the social and ethical implications of their work. While these questions implicate other areas of engineering, they are particularly salient for ML—a fast-moving field, attracting a great deal of ethical scrutiny. Thus, we put forth that ethical consideration is a core component of the practice of machine learning, and therefore, it should be a core component of machine learning education. With this in mind, this article advances the long tradition of research on computing ethics education, and considers the importance of, and potential best practices for, tackling ethics as a topic in ML education.

To begin, Section 2 provides additional background on computing ethics and motivates the importance of ML ethics. In Section 3, we consider the current state of ML-ethics education by reviewing a sample of ML course syllabi. Section 4 provides a systematic literature review of current scholarship within the ML ethics domain, with the goal of identifying the most common dilemmas and questions pertinent to ML practice. Section 5 then reports on a pilot study exploring the pedagogical utility of offering the identified questions as a framework for students learning ML. Integrating these findings into something actionable, Section 6 provides example course assignments that embed ethics within core ML topics. Finally, in Section 7, we discuss the overall findings of our work, including limitations and next steps.

2 BACKGROUND AND MOTIVATION

2.1 The Role of Ethics in Computing

Because machine learning is inextricably linked with computing, and computing has a longer history than ML, it is worth briefly reviewing what we know about ethics in computing. The potential of computing technologies to raise ethical and social issues that differ fundamentally from those raised by other technologies has been discussed since the very inception of digital computing [98]. While there are earlier examples of philosophical thought surrounding computing and ethics, such as explorations of the difference between a computer-based decision and a rationalized,

human choice [96], applied, disciplinary discourse only started in the 1980s and 1990s. During this period, computing ethics developed into a field of applied ethics [79], and dedicated courses on computer ethics were included in curricula and textbooks on the topics were written. In addition, academic conferences (e.g., Computer Ethics Philosophical Enquiry and Computers and Philosophy) and journals (e.g., Ethics and Information Technology) were created. There have also been a number of overviews of the field [9, 79] as well as anthologies aiming to cover the main topics [14, 45]. As computing technology gains complexity, so do the surrounding ethical implications. Not only must computing ethics scholars work to keep up with the movement of the field, but it also becomes important to consider how to teach in the context of emerging ethical conundrums.

2.2 Ethics Education in Computing

This need for ethics has long raised questions around how to raise awareness and interest of computing experts in the social and ethical aspects of their work, for example, by including ethics in standard curricula or professional accreditation. However, how best to approach ethics in computing education has been a longstanding question. For example, in the United States, the Accreditation Board for Engineering and Technology (ABET) requires that accredited computer science programs must produce students that have "an understanding of professional, ethical, legal, security and social issues and responsibilities." However, it does not specify how this should be accomplished, though the most common options have been required standalone ethics classes or the integration of ethics material into other required classes.

In 1996, a report from an NSF-funded project to inform computer science ethics curriculum recommended that ethics content should be integrated into core computer science classes, as a preferable solution over simply having a standalone ethics class [55]. This suggestion follows logic from other disciplines, where there is acknowledgment that ignoring ethical issues as they arise marginalizes ethics—that ethics should be seen as a necessary part of daily practice rather than a public relations digression from what is actually important [69]. In addition to research related to pedagogy for computing ethics, such as inclusion of codes of ethics [11, 46], using active learning [46] and hands-on exercises [95], there have also been efforts to integrate ethics into existing courses such as human-computer interaction [77] and introductory computer science courses [51].

2.3 The Need for Ethics in the Field of Machine Learning

Emerging ethical dilemmas are already touching the practices of computing professionals who use ML to solve problems, forcing them to make difficult decisions. In fact, the need for ethical consideration when using ML techniques has been frequently noted [34, 74].

For example, one may be asked to develop a model to predict the healthcare cost of a prospective employee by tracking and analyzing eating habits and exercise routines [40]. To address this type of project, ML engineers, and the management of that organization, need to understand a range of underlying ethical issues such as fairness (what training data should be used to ensure there is no gender bias in an ML system used to rank job applications) and privacy (is it okay to data mine "public" social media data to train models to infer personal attributes and identity). Then, the ML engineers and managers need to work together to approach those dilemmas thoughtfully.

Another example where ethics has been intensively debated is ML's application to criminal justice, specifically a recent controversy that involved a Florida county using the COMPAS recidivism prediction score to determine sentencing. This algorithm was found to have false-positive and false-negatives that created a disparate impact for African Americans [1]. While some disagreed with this view [34], recent studies deepened our understanding of trade-offs in how fairness is defined [19] and demonstrated the tension between improving public safety and satisfying the prevailing notions of algorithmic fairness [22].

These examples demonstrate the potential challenge in identifying bias and the importance of considering the rights of different stakeholder groups whose lives may be disparately impacted by an ML system. Thus, it is not surprising that Tiell and Metcalf [88] have argued that ML introduces new classes of risk to organizations. From a broader perspective, since ethics is thought to be a key component in helping to determine the acceptance of new technologies [79], it is important that ML practitioners consider the harm that might arise from their work while still allowing for the novel adoption of ML algorithms. Without exploring these questions, the unethical use of ML could impact the reputational and economic well-being of an organization, such as the public's well publicized reaction to Target's alleged prediction of a teenager's pregnancy based on the buying patterns of that teenager [2].

The conventional approach of integrating ethical considerations into a field is to leverage a code of ethics, such as the Association for Computing Machinery (ACM) code of ethics and professional conduct. In 2018, the code was updated for the first time since 1992, in part due to a recognition of "amazing changes in computing technology" and "important changes in how deeply that technology is integrated into social structures and into people's daily lives" [12]. However, recent research suggests that the currently available codes and frameworks might not be sufficient for teaching ML ethics in that the full breadth of ethical challenges that computing professionals might encounter has not yet been fully explored [50, 72, 83, 93].

2.4 Ethics Education within a Machine Learning Context

In 2014, a survey on the broader field of data science (which typically is viewed as the process of collecting, cleaning and analyzing data using a range of techniques ranging from ML to information visualization), revealed that 76% of respondents recognized the need to include ethics in data science education [66]. However, a survey of the top 15 data science programs, which was also done in 2014, showed only a handful of programs provided a course in data science ethics [57]. In 2016, a more in-depth analysis of data science programs did not even include ethics in its coding scheme [87], which could mean that ethics was not present in the programs or that it was not deemed important enough in the analysis itself. While we are not aware of similar studies specifically for ML, we expect that the need is also strong for ML. Supporting this need, there have been calls for required ethical training and participative ethical assessments when using ML in industry [50].

Recently, there has been some progress. For example, the number of standalone ethics classes in the areas of data science and artificial intelligence have been on the rise in recent years [76]. Other educational efforts specifically related to ML have also occurred outside of the traditional classroom. For example, the FATML community (Fairness, Accountability, and Transparency in Machine Learning, www.fatml.org) brings together researchers and practitioners concerned with fairness, accountability, and transparency in machine learning. While their focus is broader than ML ethics education, much of the focus is directly relevant to what one might teach ML students.

Hence, there is a need to better understand the current state of ethics in ML education in formal university classrooms, which we discuss next (Section 3).

3 CURRENT STATE OF ETHICS IN ML-RELATED COURSES

Since ML and data science programs are relatively new, many schools do not yet have specific programs in this field (particularly at the undergraduate level). For the majority of universities that do not have this type of specialized program, ML is being taught in other departments, primarily computer science. Therefore, we decided to look to these more established programs, wondering how educators who are primarily responsible for ML curricula are teaching ethics as part of their courses, if at all.

Analysis of syllabi is a common method for considering curriculum requirements in educational research [18]. It is also an imperfect measure, since syllabi do not always capture the fine-grained details of classes; however, we judged this to be an appropriate approach for an exploratory analysis to help provide a general sense of the current state of ethics in ML education.

3.1 Course Analysis Methodology

We began with the following research question: "Are existing ML-related courses in computer science including ethics as part of their coursework?" We focused on the top twenty computer science programs at U.S. universities, as ranked by their graduate programs in *U.S. News and World Report* in 2018. Though these are likely not representative of all ML, data science, or even computer science programs, we chose these since, regardless of their quality, the prestige generated from the rankings might result in other programs looking to them as a model. Specifically, our dataset included all ML-related courses (both graduate and undergraduate) in these top twenty programs.

Our data collection involved, for each of these identified universities, finding the online course catalog and identifying courses related to the following keywords: "machine learning," "data science," "big data," "data mining," and "artificial intelligence." For each course, we searched for a syllabus either available from the course listing or by searching for the course title on Google. If there was more than one version of a course syllabus available, then we used the most recent one.

We used a mixture of content analysis and iterative, open coding, which is consistent with previous research reviewing syllabi to study learning objectives, course content, or other objective measures of what is in a class [31, 85]. Specifically, for our analysis, one researcher conducted open coding of a subset of the syllabi, looking for ethics-related topics such as fairness and privacy. Multiple researchers then met to discuss the codes and the way they were being applied, and to collectively create heuristics for determining whether a course likely included ethics content.

For the purpose of this exploratory study, we categorized all the identified courses with a simple binary: yes, they appear to include some ethics-relevant content, or no, they do not. For courses where we could not find a syllabus, we simply used the course description to make this judgment. Three of the authors conferred on these codes, and we erred on the side of inclusivity if there was a subjective judgment as to whether a syllabus indicated ethics-related content. The addition of AI courses in our dataset was also an attempt at more inclusivity, since these courses may or may not include ML content. For each program, we also identified whether they had standalone general ethics courses, separate from the ML courses in our dataset. Given the limitations of our approach (such as not having all details about every class), for ethical reasons, we do not mention any specific programs or courses in our discussion of our findings.

3.2 Course Analysis Findings

Across the 20 programs, we identified 186 ML-related courses (average of 9 per university). Table 1 shows the data for each program, differentiating between two categories of ML classes: those that specifically focused on ethical/societal issues of ML (e.g., a course on AI and Society or Fairness in Machine Learning) and those that did not, which were more likely to be technical or general ML-related courses (e.g., Applied Machine Learning or Foundations of Data Science). While only 14 (7%) of all the ML courses fell into the category of ethics-specific, it is encouraging to find that ethics in ML and data science is given enough weight to carry standalone classes in nearly half of the programs in our dataset. However, we make the case in this article that the integration of ethics into technical ML classes is what will create the best learning opportunities. In addition, only 22 (12%) of technical ML courses in our dataset explicitly included some ethics-related content. Overall, we observed no explicit mention of content related to ethics in the vast majority of the courses we analyzed (150 of the total 186 courses identified).

Table 1. ML Courses with and without Ethics

School	Total ML related Courses	Ethics-Specific Courses	Technical Courses w/ Ethics	Courses without Ethics
A	9	1	4	4
B	13	4	1	8
C	9	3	0	6
D	7	0	2	5
E	11	1	2	8
F	8	1	1	6
G	13	1	2	10
H	9	0	2	7
I	11	2	0	9
J	17	0	3	14
K	8	1	0	7
L	8	0	1	7
M	8	0	1	7
N	9	0	1	8
O	9	0	1	8
P	10	0	1	9
Q	8	0	0	8
R	5	0	0	5
S	6	0	0	6
T	8	0	0	8
Total (# Courses)	**186**	**14**	**22**	**150**
Total (# Institutions)	**20**	**8**	**13**	**20**

Note that there was a wide variety across the programs. For example, three programs did not explicitly call out ethics content in any of these courses. In contrast, one program includes ethics in nearly 40% of all of their ML courses, including 4 ethics-specific courses. With respect to the content of these courses, the most common relevant topic (beyond simply mentioning "ethics") was privacy, followed by fairness. Other commonly mentioned topics were bias, transparency, accountability, and responsibility. We also saw some excellent examples of courses that were specific to these topics–for example, a course on fairness and data validity.

Because this dataset was limited to ML-related courses, it did not include general computer science ethics courses that may also include some discussion of these topics. We supplemented our analysis by identifying these courses as well–for example, a course on computing ethics and policy or on computer science professionalism. We found that the majority (65%) of these programs had such courses, but our reading of course requirements, where available, revealed that these courses are more often elective courses. We also coded for whether, based on a similar syllabi analysis, the course mentioned ML-specific topics such as bias or fairness; nearly all of the offered courses did. Therefore, students may also be getting ML ethics content in these additional standalone courses.

It is also possible that a program could encourage or even require students to take courses taught in other departments such as philosophy, though we speculate that it is less likely that these courses would offer ML-specific or even computing-specific content. Moreover, if standalone ethics courses (either those that are ML-specific or those that are general computing ethics) are typically not required, then students who may not already have an interest in ethics or do not think of ethics as part of ML would most frequently only encounter ethics in technical ML courses—which as our analysis shows, in most cases would not be available.

3.3 Course Analysis Discussion

In general, our findings echo Martin's [57] with respect to ethics content not being an integral part of this field, though we do caution against overgeneralizing our conclusions, particularly since this exploratory analysis was limited to a subset of U.S. computer science programs, primarily at research institutions. However, as a way of interrogating at least part of the current landscape

of ML ethics education, our analysis supports anecdotal evidence that standalone data science ethics courses are becoming more common [76], while also revealing less frequent integration into technical ML classes. Though the majority of programs in our dataset do include some ethics content, either in a subset of technical courses or in elective standalone courses, our analysis suggests that ethics is most often not presented as a core component of ML, data science, or AI at these universities.

There are many reasons that could account for this, ranging from an instructor not knowing what to teach or how to teach it, to an an instructor not thinking that it should be part of the material. In response to the 1996 CACM article that proposed ethics integration, one computer science professor wrote a letter to the editor arguing that ethics is "not computer science" and that it was "difficult to imagine a computer scientist teaching these things" [54]. We take the optimistic stance that this is not the dominant viewpoint today, and therefore, in this article, we tackle the problem of how best to teach ethics in the context of machine learning.

4 TOWARD A FRAMEWORK FOR ETHICAL EDUCATION IN MACHINE LEARNING

Given the recent emergence of the field, it might not be surprising that ML-related programs have yet to adopt a robust set of materials for teaching ethics. To make headway toward a foundation for such materials, we sought to identify major concepts as well as questions being asked by practitioners and researchers running into ethical dilemmas within ML. Past work has suggested that creating an ethical framework that establishes and connects key vocabulary, questions, and practices might be the requisite first step in building the needed foundation for teaching and practicing ML ethics [89, 93].

Defining a framework involves the outlining of questions, concepts, and rubrics that can be applied by students to encourage critical thinking and ethical reflection within their learning and practice. A framework would ideally bolster the use of ethical ML techniques and choices by fostering the progress of ML, while also being attentive to the protection of individual and group rights [34]. Such a framework could also support the need that ML teams have to systematically address the ethical impact and implications of their work using a consistent, holistic approach [89]. It could also help address questions concerning the responsibilities and liabilities of people in charge of ML processes, strategies, and policies.

With this need and approach in mind, we conducted a systematic literature review (SLR) of current ML-related scholarship that touches on ethics. Our goal was to curate the most common ethical terms, dilemmas, and challenges identified by contemporary experts in the ML field. The results of our literature review allow us to make recommendations about foundational ethical questions that should be included in an framework that is focused on helping students understand the potential ethical conundrums that might be encountered within an ML project.

4.1 SLR Methodology

We leveraged Kitchenham and Charters' guidelines [47] for conducting an SLR, which guided our planning, conducting, and reporting of this analysis.

4.1.1 Planning the SLR. In planning our SLR, we first composed definitions of the search space, search terms, publication period and the language of the search. Note that the search terms included "machine learning," "data science," and "big data," but not AI or Artificial Intelligence. This was due to the fact that though AI coursework often involves ML techniques, there are a myriad ethical issues in AI that are unrelated to its technological underpinnings (e.g., job loss, warfare, the possibility of Artificial General Intelligence) [52]. This omission allowed our search to be more focused on ML ethical issues that project teams need to directly confront.

We composed the search string for each database manually, restricted to articles published after 2009, since older articles would likely not capture the emerging issues and challenges in this new domain. We also restricted our search to peer-reviewed articles. While Kitchenham and Charters [47] note that other sources (e.g., ArXiV) might also be useful, they also note that this limitation serves as a proxy for checking expertise and quality. We conducted this search in April 2018.

To determine whether a paper should be included in our analysis, we defined the following inclusion criteria: (i) papers published in the ACM Digital Library, IEEE Xplore, Scopus, or the Web of Science; (ii) papers that were written in English; (iii) papers that included ("machine learning" or "data science" or "big data") and ("ethics" or "ethical"); and (iv) papers that were published after 2009. In addition, the following items comprised our exclusion criteria: (i) papers that did not meet inclusion criteria; (ii) papers that did not explicitly focus on ethics within an ML context, but rather, only referred to ML or ethics as a side topic; and (iii) papers that did not focus on ethical challenges an ML project might encounter, but rather, focused on high-level societal ethical considerations beyond the possible control of the organization supporting the data science effort.

4.1.2 Conducting the SLR. In this initial search of the identified repositories, we identified and retrieved 1,170 papers, then manually inspected titles and abstracts to apply the exclusion criteria. After this second phase, 266 papers remained. We then skimmed these papers to confirm our inclusion and exclusion criteria, and then removed duplicates. These steps resulted in a final corpus of papers.

According to the guidelines provided by Kitchenham and Charters [47], we defined a data extraction process to identify the relevant information from the 102 papers to extract: (i) review date; (ii) title; (iii) author, (iv) reference; (v) database; and (vi) year of publication. Once the extraction was completed, we continued with the guidelines provided by Kitchenham and Charters and used content analysis to explore the key ethical concepts discussed within each of the papers, recording these concepts as part of the data extraction. Specifically, we analyzed the papers through an iterative process of item surfacing, refinement, and regrouping to generate the key themes used as our framework to describe the ethical challenges noted in the papers.

Finally, we assessed the repeatability of our data extraction and categorization by using an interrater analysis among two researchers who independently coded the papers [33]. After training, the coders agreed on 89% of the coding decisions. Disagreements were discussed and agreed upon to create a final coded data set.

4.1.3 Approach to Analysis. Because our review was conducted for the purpose of creating practical, usable guidelines for students, within these key themes, the review focused on ethical challenges and key questions that ML practitioners or researchers should contemplate as they work on a project. However, what constitutes "ethical" for a given situation is rarely a concrete or simple consideration. Traditional ethical theories often serve as instructive frameworks for analyzing issues and dilemmas. Though an analysis under different frameworks may arrive at different conclusions, they are helpful in systematically thinking through dilemmas and understanding arguments made by others. In identifying questions and building our framework, we explored popular ethical theories, and focused on three described by Briggle and Mitcham [10] that cover a range of perspectives and fit well within the scope of our framework:

- **Consequentialist** theories led us to focus on the effects or consequences of alternate courses of action. It has strength in evaluating decisions with complex outcomes, in which some people benefit and some are harmed. It is weaker in situations where the consequences of an action cannot be predicted.
- **Deontological** theories helped us focus on using concepts such as duty, rights and fairness to evaluate courses of action. Duty-based ethics can be inflexible, since obligatory duties do

not leave flexibility for evaluation of the harm they might do. Rights-based ethics lead to decisions based on the rights of those affected by the decision but is less helpful in situations where rights are not impinged. Justice-based ethics focus on fairness and equality.

- **Virtue** theory stands apart, in that rather than considering a specific situation or act, it considers all actions of an individual's life and whether these collectively constitute the actions of a virtuous person. This holistic view makes it more challenging to apply to individual situations, or to consider specific motivations.

These theories have been leveraged across a range of domains, including within a software engineering ethics context [91], as well as within the broader context of science and technology [4]. Following this usage, we leveraged these theories to help shape the questions that were generated during our SLR.

4.2 SLR Findings

When we began our systematic analysis of the papers in our data set, we noted that the majority of the identified papers were published within the past few years. In fact, only four of the articles were published prior to 2014. This increase in the number of publications is not surprising, as this coincides more broadly with the increasing use of ML across a range of contexts.

We are also not the first to conduct literature reviews in this space. For example, Mittelstadt et al. [62] focused on the ethical issues of algorithmic mediation, and noted six types of ethical concerns raised by algorithms: (1) inconclusive evidence, (2) inscrutable evidence, (3) misguided evidence, (4) unfair outcomes, (5) transformative effects, and (6) traceability. In focusing on medical information and big data, Mittelstadt and Floridi [63] identified five areas of concern: (1) informed consent, (2) privacy (including anonymisation and data protection), (3) ownership, (4) epistemology and objectivity, and (5) "big data divides" created between those who have or lack the necessary resources to analyze increasingly large datasets. Others consider the broader area of big data, such as Salleh and Janczewski [71], who focused on security and privacy issues within a big data context.

Based on our analysis of the papers in our SLR, as well as leveraging these prior reviews, we identified three key areas of focus: (1) Oversight Challenges, (2) Data Challenges, and (3) Model Related Challenges. We discuss these challenges, themes, and key questions below, and we summarize them in Table 2.

4.2.1 Oversight Related Challenges: Accountability and Responsibility. Though ethics and law are two separate concepts (i.e., what is legal may not be what is ethical and vice versa), formal rules such as laws and regulations are important to consider in the context of ML projects [83]. In fact, government regulators have long enforced legal restrictions to prevent discrimination, unintended or otherwise, in industries such as financial services [26]. More recently, these laws and regulations cover concepts such as privacy via HIPAA (Health Insurance Portability and Accountability Act) in the U.S. [81] and GDPR (General Data Protection Regulation) in Europe [16].

However, regulations also tend to lag behind technology improvements [101] and the use of ML likely introduces entirely new classes of risks [88]. Therefore, since ML is a new field, many norms and regulations may not yet have been explored or defined [60, 86]. Thus is it not surprising that governments are re-conceptualizing what role the law can play in controlling and regulating the use and misuse of data and ML [16] and that one of the most important barriers and challenges to the use of ML are potential legal issues [16]. In addition, it can be a challenge to apply these rules. For example, data ownership might be ambiguous—is it the person who created the data or the technical device or company that recorded it [8]. The ownership of this type of data is unclear [42], which makes it difficult to appropriately apply laws and regulations with respect to data ownership and privacy.

Table 2. Ethical Questions About Machine Learning

Challenge	Theme	Questions
Oversight related challenges	Accountability & Responsibility	1. Which laws and regulations might be applicable to this project?
		2. How is ethical accountability being achieved?
Data Related Challenges	Data Privacy and Anonymity	3. How might the legal rights of organizations and individuals be impinged by our use of the data?
		4. How might an individuals' privacy and anonymity be impinged via aggregation and linking of the data?
	Data Availability and Validity	5. How do you know the data is ethically available for its intended use?
		6. How do you know the data valid for its intended use?
Model Related Challenges	Model and Modeler Bias	7. How have you identified and minimized any bias in the data or the model?
		8. How was any potential modeler bias identified, and then if apprioriate, mitigated?
	Model Transparency & Interpretation	9. How transparent does the model need to be and how is that transparency achieved?
		10. What are likely misinterpretations of the results and what can be done to prevent those misinterpretations?

Thus, it is important to consider which laws and regulations might affect the conduct of the project, what these laws are designed to protect or accomplish, and what the impact may be of not taking them into account. This decision-making process suggests that a team member must understand the laws and regulations that pertain to a project—e.g., the various national and local regulations that might apply to privacy, confidentiality, data protection, or intellectual property rights. Therefore, there may be cause to involve experts such as Chief Information Security Officers, general counsel, compliance officers, and where they exist, ethics officers. This suggests that to help think about the consequences of their project, each ML project team should ask:

Q1: Which laws and regulations might be applicable to this project?

In addition, there must be ethical accountability [39]. In other words, it should be clear who will be accountable for the harm that could be done by the use of this technology [62]. Accountability includes ensuring the project team proactively identifies potential stakeholders and evaluates harms such as possible disproportionate effects that may arise from the application of a model. Note that the harm might be legal in nature (e.g., bias) or it could relate to other important societal rights (e.g., self-determination, employment, health care).

However, the question of who and how one should be accountable can often be challenging to define [62]. One approach to ethical accountability is via an ethical review process [84]. For example, where each participant in the project asks critical questions about the potential impact of their contribution and has the opportunity to examine and discuss the viewpoints of participants with others [50]. A complimentary review process for ethical accountability of an ML algorithm is to have people review the algorithm, perhaps via algorithmic accountability reporting [67]. In this view, the focus is on the external behavior of the algorithm—which is how society regulates human behavior—not by looking into their brain's neural circuitry, but by observing their behavior and judging it against certain standards of conduct.

Thus, each ML project should ask:

Q2: How is ethical accountability being achieved?

4.2.2 Data Related Challenges: Privacy and Anonymity. An individual's right to choose which of their activities and facts are shared with others is an important consideration for ML and data science teams [56]. In a digital age, this includes both what the individual chooses to publish and their ability to control with whom the data is shared, including concepts covered by recent regulations such as GDPR—the European General Data Protection Regulation [53].

Privacy issues focus on who should control access to data and ownership concerns—not just who owns the collected data but which rights can be transferred and what obligations collecting or receiving such data entails [58, 97]. Further, once collected, data could be shared with third parties adhering to different privacy policies [65]. This is complicated by the additional fact that even if the users consented to the original capture and use of their data, the users do not always know when it is later used in ways they did not expect or desire [36]. Hence, project teams (and organizations) should not enter into confidentiality agreements that preclude explaining who their data partners are, as well as making the data supply chain visible so that an individual or organization has the ability to ensure no data misuse [57].

In other words, for the project to be ethical, the organization must have the right to use the data for their specific purpose. This suggests each ML project should ask:

Q3: How might the legal rights of organizations and individuals be impinged by our use of the data?

While the need for anonymity is not new to the computing field, the thought process with respect to how to ensure anonymity must be re-examined with the emergence of advanced data science linking techniques in that de-identification is one of the most challenging current ML-related privacy issues [16]. For example, it has been noted that people can be re-identified from anonymous data using zip code, birthdate and gender with 87% accuracy [40]. The impact of aggregating and linking data, and the ability for harm to arise from that information, has been noted as differentiators from other fields [82]. The cause for concern in this instance is not in the collection of data itself, which may be innocuous in isolation, but its aggregation, correlation and de-anonymization [65]. In one example, Netflix was sued by a closeted lesbian mother after researchers demonstrated that Netflix data published for a competition, when combined with data from the IMDB website, uniquely identified customers and their viewing preferences [30]. In this situation, Netflix failed to understand either that this re-identification was possible, or that this re-identification was problematic, revealing a lack of knowledge of either technical or ethical issues in their research.

Hence, consideration must be given as to how privacy will be maintained through the transmission, storage and merging of the data [64]. Consideration should also be given to whether this use of the data represents an intrusion on the privacy of the data subjects when considered from their perspective [56]. This suggests each ML project should ask:

Q4: How might individuals' privacy and anonymity be impinged via aggregation and linking of the data?

4.2.3 Data Related Challenges: Availability and Validity. Being able to access and collect data does not mean that it is ethical to use that data [7]. A very early example of data being used in a way that the data provider did not intend involved the 1940 U.S. census. Intended by law to be kept private, the census was used in 1943 to force all persons of Japanese ancestry into internment camps in the U.S. [5]. Even today, terms of service might specify the legality of collecting data, but sometimes data rights are more ambiguous or more complicated, and in reality, obtaining individuals' true informed consent is very challenging [37].

For example, even though content might be "public" (e.g., a tweet), there might be considerations beyond this accessibility for the ethical use of that content—such as, who is using it and in what way it is being used [32]. In a different example, imagine that an energy supply company finds a way to monetize its customers' smart meter data by selling it to an organization that wants to learn about how people live, yet has no intention of ever selling any product directly to those customers. The data would provide additional revenue to the energy supplier, yet there might be no incremental benefit to its customers. In this situation, it is not clear who owns the data and if that data is being misused. Perhaps the customers should expect to share some of the bounty via reduced energy pricing [38].

More generally, personal data is often used for purposes beyond its intended purpose and many users would consider such practices a violation of their right to privacy [65]. In reality, understanding if consent was given to use the data for its proposed use is still more in the grey area of feelings, opinions, and right treatment [8]. Hence, care must be taken to understand who owns the data, what are their rights and expectations, and is the data being used the way that it the person (or entity) that contributed the data intended? This suggests each ML project should ask:

Q5: How do you know that the data is ethically available for its intended use?

Data validity is another key challenge [44], and ML engineers have a responsibility to ensure that the data that they use is suitable for their needs [92]. One aspect of data validity is data accuracy, since accuracy is critical for ethical assessment and legal probity [80]. For example, imputing missing values or excluding records with missing values could have a significant impact on the downstream analytical results [7, 35].

Another data validity concern arises from "fitness of purpose" questions about how specific data is used. For example, while patient-generated health data is useful in many situations, there are situations where it might not be appropriate [21]. Specifically, due to a lack of oversight, the fitness of purpose of this data in some ML contexts could be questionable due to challenges such as missing data or accuracy of the measurements (ex. if there is no lunch reported, did the person skip lunch or just not report it) [21]. In a different example, a growing number of states use data from students' standardized test-scores to develop teacher performance scores. The output from these models is sometimes used in decisions about teacher tenure, dismissal and compensation. However, many question the accuracy of a single student test score as input into such a model [13]. It has been noted that when any one student takes a math test, on any one day, there is a huge uncertainty around that score. It could be the student got lucky this year, and guessed two or three right questions. Or the student might not have been feeling well. Consequently, the score on any one day is not necessarily a good reflection of a student's attainment level. Hence, even if the database has the correct scores stored, the data from one test might not be appropriate as a key input for the teacher evaluation model [13].

In fact, data publishers need to assist ML practitioners, so that they can properly ensure fitness of the data used for ML [41]. Thus, this theme not only covers the accuracy of the data but also whether the data is appropriate for the problem being addressed, in that ML practitioners needs to ensure "fitness of purpose" with respect to how data is used. Otherwise, data can be taken out of context, or might not be used as the data provider intended. Taken together, this suggests each ML project should ask:

Q6: How do you know that the data is valid for its intended use?

4.2.4 Model Related Challenges: Model and Modeler Bias. ML models can be built using data that records a bias, and thus, the model might also learn that bias, and as such, systematically

disadvantage a societal sub-group [24]. This can lead to ethical problems such as group discrimination (e.g., ageism, ethnicism, sexism) [34]. In other words, bias might come from the fact that the data used to build the model was biased [24].

As models are built, the ML engineer selects variables to use as features in the model. It is widely understood that it is unethical, or in many use cases illegal, to make decisions based on variables that describe protected categories such as gender, race, religion. To do so may be discrimination, and is a form of bias. In fact, ML practitioners should be aware that, even with the best of intentions, using commonly protected attributes exposes them to potential legal risk [26]. Beyond the explicitly protected categories, they must also consider other potential biases [34]. For example, using accelerometer and GPS data from smartphones, an organization predicted potholes. However, older, poorer people were less likely to have smartphones [25]. This meant that the smartphone data was missing information from significant parts of the population—often those who have the fewest resources.

More generally, analytics allows for a new type of algorithmically assembled group to be formed that does not necessarily align with classes already protected by privacy and anti-discrimination law, or addressed in fairness and discrimination-aware analytics [61]. In this situation, individuals are linked according to offline identifiers (e.g., age, ethnicity, geographical location) and shared behavioral identity tokens, allowing for predictions and decisions to be taken at a group level rather than an individual level. A simplistic example of such a group is "dog owners aged 38–40 that exercise regularly." Being identified as a member of this group could drive a variety of automated decisions with harmful or beneficial effects for individual members, such as a preferential rate for health insurance [61].

To account for this potential bias, some have started to create frameworks for identifying how discrimination may enter into the ML process [26]. While these frameworks are still evolving, at a minimum, each ML project should ask:

Q7: How have we identified and minimized any bias in the data or in the model?

Another concern is the subjectivity within the model building process [26], in that model building involves subjective decisions, and that these decisions can result in biases and prejudices [64]. In other words, there can be subjectivity when decisions must be made, such as with respect to what metric one should optimize, which algorithm to use, which data sources to use or whether one data point should be used as a proxy for a missing fact [64, 73]. For example, when the admissions department of a university designs an algorithm to determine how to allocate their annual scholarship budget, difficult choices need to be made about whether to rank applicants based on their economic background, race, prior academic achievements, or the faculty to which they apply. Each choice is reflective of different ethical and political positions, and the process of choosing between them can expose power relationships between different individuals, academic ranks, and institutional structures [100]. Furthermore, biases can also exist in the ML engineers [35]. Thus, unfortunately, with the aim of correcting for known biases, well-meaning ML practitioners may introduce new and unknown biases [100].

This is reinforced by Boyd and Crawford [6], who note that "researchers must be able to account for the biases in their interpretation of the data. To do so requires recognizing that one's identity and perspective informs one's analysis." This suggests each ML project should ask:

Q8: How was any potential modeler bias identified and then, if appropriate, mitigated?

4.2.5 Model Related Challenges: Transparency and Accuracy. Transparency is generally desired, because algorithms that are poorly predictable or explainable are difficult to control, monitor and

correct [90]. However, algorithmic outcomes of machine learning are often difficult to interpret, even by experts, and an explanation in understandable terms as to why a specific decision is recommended often cannot be supplied—which makes explainability and comprehensibility very difficult [62, 64]. Thus, many models are effectively a black box to everyone, layman and expert alike [29]. Model transparency is particularly important when model output might disadvantage a certain subgroup (or appear to disadvantage a specific subgroup), or in situations where there is a high degree of regulation or a right of challenge (e.g, lending money). In fact, many have noted serious ethical concerns can arise about the equity and fairness of an opaque, inexplicable, and potentially biased process helping to make such radically life-changing decisions [49, 62].

Neural networks, a popular ML technique, have this transparency/explainability challenge. When using a neural network, a middle layer (or more than one) is inserted connecting input and output. The weights connecting input variables to the middle variables, as well as those connecting the middle variables to the output variable, are adjusted via several iterations within model development. The end model obtained displays all those weights, but cannot be interpreted as to how much the various input variables contribute to the outcome. When transparency is needed, some have argued that the models such as neural networks should not be used, and one should use models simple enough to allow some explanation, such as explaining which covariate is driving a particular decision—perhaps even reduced to logistic regressions [38].

However, others have noted that achieving transparency via technical explanations is not the only path forward. Rather, one could also consider institutional processes, documentation, and access to those documents as a way to explain the behavior models used [75]. In other words, if justification requires understanding why the model's rules are what they are, one should seek explanations of the process behind a model's development and use, not just explanations of the model itself [75]. This approach helps to address a different challenge with respect to transparency—that some algorithms are proprietary and need to be kept secret for the sake of competitive advantage, national security or privacy [62]. Thus, while there is still ongoing debate with respect to how to achieve model transparency, at a minimum, one should ask:

Q9: How transparent does the model need to be and how is that transparency achieved?

Furthermore, most predictive models are statistical in nature. They provide no guarantees; rather, they tell us about areas where an increased probability of an outcome might guide us to act differently [26]. For example, one seldom finds a classifier with perfect or even near-perfect predictions. [26]. With this in mind, the ML engineer must ensure that the analytical decision reflects the scale, accuracy and precision of the data that was used in creating the model [20], and that the results and conclusions should be presented along with information on the range of circumstances for model validity.

Due to this, an ML practitioner's ethical responsibilities do not end with the completion of a model. They also have a duty to explain the implications and limitations of using a model and it is crucial that those who devise the analytics clearly understand and explain their impact [35]. This suggests each ML project should ask:

Q10: What are likely misinterpretations of the results and what can be done to prevent those misinterpretations?

5 PILOT STUDY ON USING THE IDENTIFIED ETHICS QUESTIONS

To evaluate the potential usefulness of the ethical questions identified via our SLR, we conducted a pilot study that explored whether students could understand the questions and, more importantly, if students could use the questions to more easily identify ethical situations within an ML

project context. Our aim was to better understand if the questions were helpful in eliciting an ethical thought process, and if so, integrate these ethical questions within our ML ethics modules, described in the next section.

5.1 Data Collection and Analysis

For this pilot study, students were given a homework assignment to analyze two publicly described ML projects and then identify the top three ethical issues within each of the two assigned ML projects. The students had two weeks to complete the assignment and were allowed to use any external resources they deemed appropriate (and cited). In total, 85 graduate students in an Introduction to Data Science class participated in the study. While most of the students were information system students, approximately 15% were business or public policy students and approximately 60% of the students had previous information-technology-related work experience. Finally, students in the class had a broad spectrum of undergraduate majors, such as information technology, computer science, engineering, and business.

Prior to the assignment, instructors led a general discussion of ML ethics and described an example ML project, which provided a vehicle to note ethical situations within that discussion. The students were also provided the list of ethical questions derived from the SLR as well as a one paragraph explanation for each of the questions. Students were told that they were free to use these questions to help identify potential ethical issues, but were not required to use the questions. Each student chose two projects from a pool of projects developed in a previous semester of the course. Once a project was selected from the pool, it was made unavailable to other students. Thus, each student analyzed a unique (to that semester) set of two projects.

We subsequently analyzed student submissions for these two assigned projects. Each student response was reviewed by two independent coders. They determined whether each issue identified fell into one of three categories: (1) non-ethical issue (issues more technical in nature, such as the appropriate use of a specific algorithm); (2) societal ethical issue (ethical issues beyond the scope of the project, such as the impact of AI on the workforce); or (3) project-relevant ethical issue. Coders agreed on 88% of coding decisions. Disagreements were discussed and resolved into a final set of coded responses.

5.2 Pilot Results

Students were able to identify 2.9 ethical issues per project, with students leveraging the supplied ethical questions to identify ethical issues directly related to the ML project for 95% of the issues identified by the students.

In terms of the other 5%, 3% of the responses discussed ethical issues that were societal in nature—issues that could not be practically addressed within the scope of the ML project. For example, one student raised the question "Would robots be citizens?," which is an interesting ethical question, but not one that could be addressed by the ML project team. Non-ethical issues comprised the other 2% of the responses (for example, questioning the technical quality of data encryption).

Note that in a previous semester, the same assignment was used, but without providing the students with questions derived from the SLR. In that semester, students identified, on average 2.2 issues, with 77% of those issues being ethical issues directly related to the ML project.

5.3 Implications

The results of our pilot study confirmed that students were able to easily understand and use the ethical questions as a starting point to explore possible ethical situations within machine learning projects. Thus, these results suggest that students can easily leverage an explicit set of questions

that helps them contemplate ethical situations. Specifically, the framework of possible questions helped students think about and identify possible ethical issues that were directly related to the project they were analyzing. This framework of questions also helped students focus on ethical issues that are actionable by members of an ML project team, and not those that are societal in nature. In summary, the framework of questions helped students analyze the projects, and hence, can be a key aspect of how to integrate ethics within an ML module, three examples of which are discussed in the next section.

6 EXAMPLE COURSE MODULES FOR ML ETHICS EDUCATION

The pilot demonstrated the value of predisposing students to an inquiry framework for considering ethics in ML projects. We will now move on to look at how ethics may be imbued into assignments that are already common to early ML classes, therefore also providing means for ethics integration even in introductory ML classes, while not requiring additional assignments. While we do not evaluate these exercises, they offer an introduction for how ethical thinking can be appropriately injected into pre-existing courses. Future work would be to research effectiveness of specific formats and longitudinal interventions; though we hope we are demonstrating to the relative ease with which ethics could be adopted all across the ML curriculum. Evidence supports integrating ethics consistently throughout the CS major [55]; thus, these examples should support how this might occur without having to completely redesign curricula.

In what follows, we discuss three example class modules where ethical questions are integrated directly into an ML assignment. Each example highlights some of the ethical questions referred to in Table 2 that could be addressed within that module. Note that these examples were selected to be representative of the type of assignments typically found within ML courses. Our formula is roughly to create a frame by doing some leading analyses that provide technical results and then relate those results to reflection questions that convey potential ethical concerns. Our examples were adapted after reading open ML course materials by UW, NYU, MIT, and CMU and noticing the common themes of teaching logistic regression, random forest classifier, and deep convolutional neural models. Below you will find direct links to the full assignment documents, which further link to Python notebooks containing walk-through solutions.[1]

6.1 Logistic Regression Module

The first example integrates ethics within a logistic regression assignment. Logistic regression is a fundamental method used for prediction, and has been used since the late 1950s [23]. It remains one of the most common ML applications and was taught in nearly all introductory machine learning courses reviewed in our syllabi analysis. In our example, we use a Yelp review dataset, and students are asked to implement a logistic regression model that predicts an entity's rating, in stars, given the text of the rating. In general, this module could easily be adapted to any assignment or lecture where a classification task is being taught within an ML context.

The assignment starts with basic tasks that one typically would do for logistic regression, such as generating histograms, plotting a confusion matrix of the model, and inspecting predicted ratings against true ratings. While the initial assignment unfolds as expected, we emphasize some practical points of reflection that get the student thinking about their choices. For instance, the student is asked to justify why they chose to split their test and training sets with a particular allocation. We further have them reflect on the causes of their inaccuracies. This may seem trivial, but it builds up the expectation that the student can, and should, justify their choices, which will support them

[1] http://github.com/ProbableModels/acm_paper_2018.

in the later ethical questions. Beyond these typical tasks, the assignment also poses a set of ethical questions that students need to address, including:

(1) "*Having summarized your data, what questions might you ask to decide if the dataset you have chosen is appropriate for the model you hope to create?*" and "*Was our training data collected in a way that is appropriate for the prediction task? Why or why not?*" In other words, how do we know the data is valid for its intended use (i.e., Q6)?
(2) "*Using a histogram or other summary statistics, did you notice a class imbalance in your training data set?*" Digging more deeply into that question, we ask, "*Why may there be such a class imbalance? Do you think this is caused by sampling or systematic problems?*" These questions help explore if there is bias in the data being used to create the model (i.e., Q7)?
(3) *"What is the most likely misinterpretation of the results?*"; "*What are the dangers of such a misinterpretation?*"; and "*Given this danger, are there restrictions one should place on the use of the derived model within a specific context?*" These questions help explore model misinterpretation and the impact of misinterpretation (i.e., Q10).

This set of ethical questions allows students to attach ethical inquiry directly to the material they are being taught and the choices they are making within the assignment. There are no right or wrong answers, but it is critical that students are able to think about these questions in terms of practical results. For example, in Questions (2) a student might point to a bias of few low ratings. A very different bias might be the fact that certain classes of people do not use Yelp, and hence, their views might not be represented in the analysis.

Finally, the assignment asks, as a thought experiment, if instead of predicting Yelp stars, the data is used to support developing a new model for predicting creditworthiness scores (i.e., a rating for whether or not a bank should give a business a loan). How might one re-adapt the methodologies previously done? That is, what might it look like to use Yelp reviews as a component of evaluating business success, and thus, credit worthiness? Thinking through this added layer can develop important contrasts providing teaching moments, given that a credit scoring model could harm individuals (e.g., denying a loan). Thus, additional questions become relevant such as which laws and regulations might be applicable for this project (i.e., Q1)? Or perhaps, is it valid to use Yelp data at all for credit scoring and what complimentary data may be required (i.e., Q6)?

6.2 Random Forest Classifier Module

The second example has students integrate ethics within a random forest classifier assignment. In this case, the model does sentiment analysis, and students use the model on the Sentiment140 dataset, which is a publicly available and commonly used dataset of positive and negative tweets, as well as the Claritin Twitter dataset, which contains tweets about the drug Claritin. Using the Claritin Twitter data offers immediate opportunities for ethical conundrums due to it containing information on clinical medical trials and personal information such as gender. Just as with the previous module, this example could easily be adapted to any classification task being taught within an ML context.

Specifically, the assignment starts with basic tasks that one might typically do for sentiment analysis, such as building a random forest classifier to determine if a tweet is positive or negative. Then, using the Claritin Twitter data, the model is used to predict if the tweet relating to Claritin has a positive or negative sentiment. Students are guided to explore the accuracy of the model, in general, as well as how the model performs with respect to the gender of the person tweeting. Again, the normal ML part of the assignment pushes toward justifying and explaining results.

The assignment ends with an additional set of questions focusing ethics. Importantly, these questions could be re-applied to most classification tasks and ideally would emphasize the importance of these thought processes. Example questions include:

(1) "*When using data that contains medical information, what laws might be applicable to building your application?*" (i.e., Q1)
(2) "*What questions might you ask about the validity of comparing the two datasets to each other?*" and "*What properties of the datasets would you look for to decide whether the two data were comparable or not?*" In other words, questions focusing if the data was valid for its intended use (i.e., Q6).
(3) "*Comparing how the model performed before and after balancing and across classes, can we conclude there was any bias in the dataset?*" and the follow-up question "*What evidence might one use to back the conclusion for or against bias?*" (i.e., Q7).

Finally, the assignment asks a broader question: Should a model like this be used to help doctors make recommendations on which patients should use Claritin? Within this context, we ask "*Would it would be appropriate to let a model like the one proposed operate autonomously, making suggestions directly to a doctor?"* Immediately, we hope to see students discuss issues related to proper oversight (i.e., Q2). Furthermore, we ask "*Is the data and model relevant for this task?*," which looks to explore if students are thinking about the ethics and validity of using Twitter data (i.e., Q5 and Q6).

6.3 Multi-model Module

The final example helps students develop methods for choosing the best ML model for a given situation by having them compare the accuracy of different ML Models applied to the same dataset. The assignment is agnostic in terms of which models are actually implemented. Rather, the goal is to provide some sense of how to analyze performance and model comparisons to understand the ramifications of choosing a particular model. Thus, the assignment could be adopted within nearly any ML class where multiple models are taught or blackbox analysis is being done.

The assignment has students using two very common datasets—IRIS and MNIST. We believe that even when working with these common data sets, it is important to emphasize the mechanics of ethical thinking so as not to suggest ethics only matters when using an obviously politically charged dataset like Stop and Frisk. Using IRIS, the student is initially asked to fit a logistic regression and a random forest classifier where the prediction target is the species of Iris plant. Then the assignment has students use the MNIST dataset, which is a dataset of handwritten digits, to try and predict the number in an image by fitting a random forest classifier and a simple convolutional neural network (CNN). The student then goes back and fits the CNN on the IRIS data. Throughout the process, the assignment guides students through a comparison of the models, for example, by having students output a boxplot that highlights the accuracy of each of the models. They are asked to explain why performance may be differing between models and datasets.

The assignment then turns to some possible ethical considerations. One question dives deeply into the comparison of the models: "*Now considering all three models, can you conclude that one model is always better than the others?*" This leads into a series of follow-ups that helps dig into why explainability might be important in certain circumstances. Thus, we ask, "*Between a CNN and a logistic regression, is it easier to explain the inner-workings of one versus the other? Try to explain how each model is working on the IRIS dataset.*" This further goes to, "*Discuss a model where being able to explain the results would be very important.*" Collectively, these questions focus on model transparency and the potential need for transparency (i.e., Q9)

In addition, students are asked to look at decision surface visualizations for a variety ML models to see how different assumptions are baked into choosing a model. We then ask, "*Choose two of*

the models, what assumptions does the model makes about data that it does not have?" Here we are pushing them to consider their assumptions and potential model bias (i.e., Q7).

The assignment ends with a broader question using a scenario. We set the scenario to assume the student is using Facebook Profile data to train a model to decide which set of products and events to advertise. The student trains three different models and notices that accuracy is, on average, the highest with a logistic regression, but sees that when employment details are known for a given user, a feed-forward neural network is even more accurate. We then specifically ask, "*Your boss asks you to consider scraping LinkedIn data to link profiles and retrain a feed-forward network. What questions might you ask about this process?*," which introduces a range of ethical considerations (i.e., Q1, Q2, and Q5). Here we are targeting real-world, job-relevant considerations that may bear on the student's understanding of the relevant ethical dimensions of this specific ML application. The module aims to prepare students to be able to meaningfully wrestle with this type of dilemma.

7 DISCUSSION

Through our analyses and findings, we hope to have raised awareness and provided direction toward ethics receiving more attention within ML curricula. Between recent scholarship, journalism, and precipitous advancements in the field, we believe the need for ethics is well-established and understood. Inversely, we recognize that traditional training for those involved in ML—e.g., computer scientists, engineers, and statisticians—does not necessarily include the training relevant to answer these ethical questions. Crucially, this training often focuses on the many difficult technical skills necessary to employ these complex methods and prove their reliability. Therefore, our goal is to lower the barrier for adopting more ethics content into existing ML curriculum while still ensuring the robustness and effectiveness of core content.

7.1 Building a Foundation

A critical first barrier is that while scholarship around ethics in ML continues to build, there are few syntheses of these materials available. For an instructor wanting to explore this field, they may not know where to start, and simple searches for common issues such as "ML ethics" or "privacy" might lead them down interesting paths, but would not provide a holistic account of the major contemporary issues, controversies, and, most importantly, pedagogical content needed to teach ethics to future ML practitioners. Thus, we see an auxiliary contribution of this work as the building of a lineage of the topic leading up to the pressing needs in ML, providing a central resource for key references relevant to teachers and researchers in this domain. Of course, many scholars are doing excellent work in the area of ML ethics, many of whom we have referenced in this article and whose work informs our framework. Despite the fact that we have likely not been able to provide a fully comprehensive view of this domain, we hope that our analysis and citations provide the necessary context to help entrants into ethical thinking.

We also hope that our exploratory analysis of the state of ethics content within existing ML courses might further motivate instructors. Though there are limitations to our syllabi review, notably the restriction of our analysis to publicly available information and to a subset of programs within the United States, as a first step, our findings both seed hope and highlight need. Overall, we see a relatively small proportion of core, technical ML classes that explicitly feature ethics content, and this is where we believe that adoption is most important. However, there is also evidence of movement in this area, with a number of standalone ethics classes offered and mentions of topics like privacy and fairness in other classes. However, it is still the case that many students may never explicitly come across ethics during the course of their ML education. When they do, it is also most likely to come at them in a single course, rather than as a patterned way of thinking inculcated into their mindset when designing and deploying ML systems.

Though we are still far from the vision of ethics thinking being baked into student mindsets, we have attempted to offer evidence that there are common and identifiable questions at the heart of this topic. The systematic literature review of ethics papers was meant to further centralize educators thinking about the topic. Going a step further than a bank of citations and a syllabi review, we hoped to offer knowledge of what the field is saying. That is, ML ethics might be discussed in many desultory write-ups, but there is largely a set of questions regularly being asked. Though this literature review may have missed some papers, the questions abstracted from the review did deliver a succinct account of the major questions being raised from a practical standpoint.

Some of our questions about the relevant laws and policies fit well with what would normally be covered in a professional ethics class. However, other questions uncovered about model validity and modeler bias get deeper into the ML practice. And it is this relationship of ethics to ML practice that we sought to build in this particular paper. The pilot study was a first step in the direction to show that ethical questions are close in proximity of real-world ML applications, if the awareness is in place. Our study is still extremely preliminary; though, hopefully it paves the ground for other educators to continue on, as it does present promise that providing the right materials and framing does benefit students in identifying these ethical challenges.

7.2 Contextualizing ML Ethics

Since prior research has shown that when ethics education is delivered in one-off courses, students do not necessarily believe the material holds importance to their career [15, 28, 78], another major thrust of our article is to provide the necessary materials to start demonstrating how ethical thinking could easily be integrated into the normal, expected course content and have that be a constant in an ML education. This was especially important in light of our syllabi review, which suggested that many programs were focused on one-off courses. Our three modules are meant to deliver this sense: that when teaching different models, evaluation approaches, or even basic statistical knowledge, ethical reflection can be brought to bare on the knowledge *in situ* without having to jump to a separate ethics lesson. We have adapted our ethical questions directly into common ML course materials such as learning about logistic regression, random forest classifiers, or convolutional neural networks. The goal is to empower a student to see that the technical choices they make—the analyses they choose to do or not do, the algorithms they use, the datasets they choose—do have human consequences and should involve ethical thinking.

7.3 Limitations and Next Steps

During our syllabi review, we may have easily missed instances where ethics was taught. This could have been due to not having access to the full course materials, an instructor discussing ethics without putting it in their syllabus, or missing a course due to the particular keywords and databases we searched. In addition, the analysis of additional universities might change the relevant statistics.

The SLR could have included additional, non-peer reviewed, content (e.g., ArXiV, blog posts). Furthermore, future research could explore the framework, and the associated key questions, using multiple ethical theories. This could focus on, for example, exploring the difference between the duty, consequence and virtue perspectives previously discussed in Section 4, or exploring other perspectives that could also prove interesting, such as Rawls' theory of justice perspective [68]. Collectively, this additional work could help enable the framework to be a robust and enduring framework, eventually of use to both educators and practitioners. Fully developed, such a framework could have additional questions and more complexity, exposing hierarchies or relationships among ethical questions, but, especially for students, this needs to be balanced with a framework

that is general and easily understood. In addition, our framework and associated questions were identified via the SLR, and thus may be specific to our current time frame, in which, for example, privacy is a hot topic. In the future, privacy may be eclipsed by another topic (e.g., political advertising and voting patterns). Hence, as the literature changes, the current framework, since it is based on the SLR, could become obsolete. Finally, since models are built from data, our framework does not currently draw a clear divide between data and ML ethics. Nor does it take a stance on engineer responsibility or algorithmic accountability standards.

In terms of creating course modules that integrate ethics into the ML classroom, there might be other core ML topics where the integration of ethical concepts is more challenging than what was encountered in our three examples. We recognize there may be some assignments where asking for deeper reflection and explanation, beyond the practical choices, is not be appropriate – such as when a student is making a first pass on a concept and barely knows the methodology. Hence, future research could explore a variety of approaches and the utility of the modules within the classroom environment, perhaps based on if the course is an introductory or advanced course.

Finally, our article does not address what a faculty's background and preparation should be to teach ethics in an ML course. Future research might explore if instructors are worried that they do not have the appropriate knowledge to address these ethical topics, and if so, how to address these concerns.

8 CONCLUSION

In this article, we offered context, insights, and practical suggestions regarding cultivating ethics as component of machine learning (ML) education. ML is a rapidly advancing field with new applications being deployed on a near daily basis. ML models are often trained on data sourced from human behaviors and their autonomous deployment leads to ML systems being a touchstone of many online interactions. Thus, the social ramifications of ML are increasing and already we are witnessing a pressing need for ML engineers to have a better understanding of ethics and social impacts of their work.

We take a stance that ethics should be infused throughout the ML curriculum, and not provided as a one-off course. The article starts by exploring the need for ethics in data science and ML. Using this backdrop as context, we look at the current state of ethics education in ML through a syllabi review from top graduate programs. Our findings indicate there is still a majority of programs where students can graduate without any exposure to ethics within an ML context. Where it is available, it is often a one-off non-required course. The most common themes being taught currently are privacy and fairness.

We moved on to surveying the current scholarship on ethics in ML and data science to offer an early foundation toward a framework of ethical topics and questions applicable to ML practice. Using this framework, we then provided three novel ML course modules that leverage the ethical questions to encourage ethical reflection as part of teaching core ML content.

We hope this work encourages and supports the increased adoption of these topics within the ML curriculum. Our assignments are simply the first phase of the vast repository of topics that would need to be attended to from an ethical standpoint within ML education. Though, the ease at which many of the lessons and activities touch upon a relevant ethical question gives hope that the burden may not be all that heavy. Continuing to improve our core frameworks for teaching ML ethics and revisiting the core content of the discipline should slowly dissipate the barriers to entry for adding some ethical thinking into one's ML course. Finally, we hope that by pooling and surveying the citations, questions, and current curricula relevant to ML ethics, we have helped future work move more effectively.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. *ProPublica* 23 (May 2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[2] Ida Asadi Someh, Christoph F. Breidbach, Michael Davern, and Graeme Shanks. 2016. Ethical implications of big data analytics. In *ECIS*. Research-in.

[3] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, 4349–4357.

[4] Sheila Bonde and Paul Firenze. [n.d.]. A Framework for Making Ethical Decisions | Science and Technology Studies. Retrieved from https://www.brown.edu/academics/science-and-technology-studies/framework-making-ethical-decisions.

[5] Grady Booch. 2014. The human and ethical aspects of big data. *IEEE Softw.* 31, 1 (2014). Retrieved from https://ieeexplore.ieee.org/abstract/document/6750430.

[6] Danah Boyd and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Info. Commun. Soc.* 15, 5 (June 2012), 662–679. DOI: https://doi.org/10.1080/1369118X.2012.678878

[7] Danah Boyd, Karen Levy, and Alice Marwick. 2014. The networked nature of algorithmic discrimination. *Data Discrim. Collect. Essays*. Open Technology Institute.

[8] Andreas Braun and Gemma Garriga. 2018. *Consumer Journey Analytics in the Context of Data Privacy and Ethics*. Springer, Berlin, 663–674. DOI: https://doi.org/10.1007/978-3-662-49275-8_59

[9] Philip Brey and Johnny Hartz Søraker. 2009. Philosophy of computing and information technology. In *Philosophy of Technology and Engineering Sciences*. 1341–1407.

[10] Adam Briggle and Carl Mitcham. 2012. *Ethics and Science: An Introduction*. Cambridge University Press.

[11] Bo Brinkman and Keith W. Miller. 2017. The code of ethics quiz show. In *Proceedings of the ACM SIGCSE Technical Symposium on Computer Science Education (SIGCSE'17)*. ACM, New York, NY, 679–680.

[12] Bo Brinkman and Keith W. Miller. 2017. The code of ethics quiz show. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education (SIGCSE'17)*. ACM, New York, NY, 679–680. DOI: https://doi.org/10.1145/3017680.3017803

[13] Sarah Butrymowicz and Sarah Garland. 2012. How New York City's Value-added Model Compares to What Other Districts, States are Doing—The Hechinger Report. Retrieved from http://hechingerreport.org/how-new-york-citys-value-added-model-compares-to-what-other-districts-states-are-doing/.

[14] Terrell Ward Bynum and Simon Rogerson (Eds.). 2003. *Computer Ethics and Professional Responsibility* (1st ed.). Wiley-Blackwell, Malden, MA.

[15] Mary Elaine Califf and Mary Goodwin. 2005. Effective incorporation of ethics into courses that focus on programming. In *Proceedings of the 36th SIGCSE Technical Symposium on Computer Science Education (SIGCSE'05)*. ACM, New York, NY, 347–351. DOI: https://doi.org/10.1145/1047344.1047464

[16] Pompeu Casanovas, Louis De Koker, Danuta Mendelson, and David Watts. 2017. Regulation of big data: Perspectives on strategy, policy, law and privacy. *Health Technol.* 7, 4 (Dec. 2017), 335–349. DOI: https://doi.org/10.1007/s12553-017-0190-6

[17] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arxiv:cs.CL/1406.1078

[18] Felicia Chong. 2016. The pedagogy of usability: An analysis of technical communication textbooks, anthologies, and course syllabi and descriptions. *Techn. Commun. Quart.* 25, 1 (2016), 12–28.

[19] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (June 2017), 153–163.

[20] Roger Clarke. 2016. Big data, big risks: Big data, big risks. *Info. Syst. J.* 26, 1 (Jan. 2016), 77–90. DOI: https://doi.org/10.1111/isj.12088

[21] James Codella, Chohreh Partovian, Hung-Yang Chang, and Ching-Hua Chen. [n.d.]. Data quality challenges for person-generated health and wellness data. *IBM J. Res. Dev.* 62, 1.

[22] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'17)*. ACM, New York, NY, 797–806.

[23] David Roxbee Cox. 1958. The regression analysis of binary sequences. *J. R. Stat. Soc. Series B Stat. Methodol.* 20, 2 (1958), 215–242.

[24] Kate Crawford and Jason Schultz. 2014. Big data and due process: Toward a framework to redress predictive privacy harms. *Boston Coll. Law Rev.* 55 (2014), 93.

[25] Susan P. Crawford and Dana Walters. 2013. *Citizen-Centered Governance: The Mayor's Office of New Urban Mechanics and the Evolution of CRM in Boston*. Technical Report ID 2307158. Rochester, NY. Retrieved from https://papers.ssrn.com/abstract=2307158.

[26] Brian d'Alessandro, Cathy O'Neil, and Tom LaGatta. 2017. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big Data* 5, 2 (June 2017), 120–134. DOI: https://doi.org/10.1089/big.2016.0048

[27] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated experiments on ad privacy settings. *Proc. Privacy Enhanc. Technol.* 2015, 1 (Jan. 2015), 65.

[28] Janet Davis and Henry M. Walker. 2011. Incorporating social issues of computing in a small, liberal arts college: A case study. In *Proceedings of the 42nd ACM Technical Symposium on Computer Science Education (SIGCSE'11)*. ACM, New York, NY, 69–74. DOI: https://doi.org/10.1145/1953163.1953186

[29] Paul B. de Laat. 2017. Big data and algorithmic decision-making: Can transparency restore accountability? *ACM SIGCAS Comput. Soc.* 47, 3 (Sept. 2017), 39–53. DOI: https://doi.org/10.1145/3144592.3144597

[30] Marina Drosou, H. V. Jagadish, Evaggelia Pitoura, and Julia Stoyanovich. 2017. Diversity in big data: A review. *Big Data* 5, 2 (June 2017), 73–84. DOI: https://doi.org/10.1089/big.2016.0054

[31] James M. DuBois and Jill Burkemper. 2002. Ethics education in US medical schools: A study of syllabi. *Academ. Med.* 77, 5 (2002), 432–437.

[32] Casey Fiesler and Nicholas Proferes. 2018. "Participant" perceptions of Twitter research ethics. *Soc. Media Soc.* 4, 1 (2018).

[33] Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Paik. 1981. Determining sample sizes needed to detect a difference between two proportions. *Statist. Methods Rates Proport.* 2 (1981), 33–49.

[34] Luciano Floridi and Mariarosaria Taddeo. 2016. What is data ethics? *Philos. Trans. Roy. Soc. A: Math. Phys. Engineer. Sci.* 374, 2083 (Dec. 2016), 20160360. DOI: https://doi.org/10.1098/rsta.2016.0360

[35] Michael Fuller. 2017. Big data, ethics and religion: New questions from a new science. *Religions* 8, 5 (May 2017), 88. DOI: https://doi.org/10.3390/rel8050088

[36] Elizabeth Goodman. 2014. Design and ethics in the era of big data. *Interactions* 21, 3 (May 2014), 22–24. DOI: https://doi.org/10.1145/2598902

[37] Daniel Goroff, Jules Polonetsky, and Omer Tene. 2018. Privacy protective research: Facilitating ethically responsible access to administrative data. *Ann. Amer. Acad. Polit. Soc. Sci.* 675, 1 (Jan. 2018), 46–66. DOI: https://doi.org/10.1177/0002716217742605

[38] Peter Grindrod. 2016. Beyond privacy and exposure: Ethical issues within citizen-facing analytics. *Philos. Trans. Roy. Soc. A: Math. Phys. Engineer. Sci.* 374, 2083 (Dec. 2016), 20160132.

[39] Peiqing Guan and Wei Zhou. 2017. *Business Analytics Generated Data Brokerage: Law, Ethical, and Social Issues*. Springer, Cham, 167–175. DOI: https://doi.org/10.1007/978-3-319-65548-2_13

[40] Andra Gumbus and Frances Grodzinsky. 2016. Era of big data: Danger of descrimination. *SIGCAS Comput. Soc.* 45, 3 (Jan. 2016), 118–125. DOI: https://doi.org/10.1145/2874239.2874256

[41] Frank Ole Hanssen, Tor Gravrak Heggberget, Jesper Bladt, Dag Terje Filip Endresen, Martin Forsius, Gudmundur A. Gudmundsson, Ulf Gardenfors, Starri Heiomarsson, Oscar Kindvall, Wouter Koch, Hanna Koivula, Eija-Leena Laiho, Matthias Obst, Flemming Skov, Anders Telenius, Nils Valland, Pawel Wasowicz, and Anna Maria Wremp. 2014. Nordic LifeWatch cooperation, final report: A joint initiative from Denmark, Iceland, Finland, Norway and Sweden. Retrieved from https://www.duo.uio.no/handle/10852/50260.

[42] Adam Harkens. 2016. "Rear window ethics" and discrimination: The darker side of big data. In *Proceedings of the European Conference on e-Government*. 267.

[43] Tristan Harris. 2016. How Technology is Hijacking Your Mind—From a Former Insider. Retrieved from https://goo.gl/GQJCE9.

[44] Clement Iphar. 2017. *Formalisation of a Data Analysis Environment Based on Anomaly Detection for Risk Assessment: Application to Maritime Domain Awareness*. Ph.D. Dissertation. PSL Research University. Retrieved from https://pastel.archives-ouvertes.fr/tel-01783958/document.

[45] Deborah G. Johnson and Helen Nissenbaum. 1995. *Computers, Ethics and Social Values* (1st ed.). Pearson, Englewood Cliffs, NJ.

[46] Michael S. Kirkpatrick and Dee Weikle. 2018. Active learning strategies for integrating the ACM code of ethics into CS courses: (Abstract only). In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education (SIGCSE'18)*. ACM, New York, NY, 1062–1062.

[47] Barbara Kitchenham and Stuart Charters. 2007. Guidelines for performing systematic literature reviews in software engineering. Retrieved from citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.117.471.

[48] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proc. Natl. Acad. Sci. U.S.A.* 110, 15 (Apr. 2013), 5802–5805. DOI: https://doi.org/10.1073/pnas.1218772110

[49] James Larus, Chris Hankin, Siri Granum Carson, Markus Christen, Silvia Crafa, Oliver Grau, Claude Kirchner, Bran Knowles, Andrew McGettrick, Damian Andrew Tamburri, and Hannes Werthner. 2018. *When Computers Decide: European Recommendations on Machine-Learned Automated Decision Making*. Technical Report. New York, NY.

[50] Sabina Leonelli. 2016. Locating ethics in data science: Responsibility and accountability in global and distributed knowledge production systems. *Philos. Trans. Roy. Soc. A: Math. Phys. Engineer. Sci.* 374, 2083 (Dec. 2016), 20160122. DOI: https://doi.org/10.1098/rsta.2016.0122

[51] Justin Li. 2017. Weaving diversity and inclusion into CS content (abstract only). In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education (SIGCSE'17)*. ACM, New York, NY, 726–726.

[52] Patrick Lin, Keith Abney, and George A. Bekey. 2011. *Robot Ethics: The Ethical and Social Implications of Robotics*. MIT press.

[53] Jenna Lindqvist. 2018. New challenges to personal data processing agreements: Is the GDPR fit to deal with contract, accountability and liability in a world of the Internet of Things? *Int. J. Law Info. Technol.* 26, 1 (Mar. 2018), 45–63. DOI: https://doi.org/10.1093/ijlit/eax024

[54] C. Dianne Martin. 1997. The case for integrating ethical and social impact into the computer science curriculum. In *ITCSE Work. Group Rep. Suppl. Proc.* (1997). ACM, 114–120.

[55] C. Dianne Martin, Chuck Huff, Donald Gotterbarn, and Keith Miller. 1996. Implementing a tenth strand in the CS curriculum. *Commun. ACM* 39, 12 (1996), 75–84. DOI: https://doi.org/10.1145/240483.240499

[56] Kelly D. Martin and Patrick E. Murphy. 2017. The role of data privacy in marketing. *J. Acad. Market. Sci.* 45, 2 (Mar. 2017), 135–155. DOI: https://doi.org/10.1007/s11747-016-0495-4

[57] Kirsten E. Martin. 2015. *Ethical Issues in the Big Data Industry*. Technical Report ID 2598956. Rochester, NY. https://papers.ssrn.com/abstract=2598956

[58] Richard Mateosian. 2013. Ethics of big data. *IEEE Micro* 33, 2 (Mar. 2013), 60–61. DOI: https://doi.org/10.1109/MM.2013.35

[59] Jacob Metcalf and Casey Fiesler. 2018. The Cambridge Analytica Scandal Shows Facebook Needs to Give Researchers More Access, Not Less. Retrieved from https://slate.com/technology/2018/03/cambridge-analytica-demonstrates-that-facebook-needs-to-give-researchers-more-access.html.

[60] Jacob Metcalf, Emily F. Keller, and Danah Boyd. 2016. Perspectives on big data, ethics, and society. *Council Big Data Ethics Soc.* Retrieved on July 12, 2019 from https://bdes.datasociety.net/council-output/perspectives-on-big-data-ethics-and-society/.

[61] Brent Mittelstadt. 2017. From individual to group privacy in big data analytics. *Philos. Technol.* 30, 4 (Dec. 2017), 475–494. DOI: https://doi.org/10.1007/s13347-017-0253-7

[62] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data Soc.* 3, 2 (Dec. 2016), 205395171667967.

[63] Brent Daniel Mittelstadt and Luciano Floridi. 2016. The ethics of big data: Current and foreseeable issues in biomedical contexts. *Sci. Engineer. Ethics* 22, 2 (Apr. 2016), 303–341. DOI: https://doi.org/10.1007/s11948-015-9652-2

[64] Stephen J. Mooney and Vikas Pejaver. 2018. Big data in public health: Terminology, machine learning, and privacy. *Annu. Rev. Public Health* 39, 1 (2018), 95–112. DOI: https://doi.org/10.1146/annurev-publhealth-040617-014208

[65] Mario Pascalev. 2017. Privacy exchanges: Restoring consent in privacy self-management. *Ethics Info. Technol.* 19, 1 (Mar. 2017), 39–48. DOI: https://doi.org/10.1007/s10676-016-9410-4

[66] Gregory Piatetsky. 2014. CRISP-DM, Still the Top Methodology for Analytics, Data Mining, or Data Science Projects. Retrieved from https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html.

[67] Iyad Rahwan. 2018. Society-in-the-loop: Programming the algorithmic social contract. *Ethics Info. Technol.* 20, 1 (Mar. 2018), 5–14. DOI: https://doi.org/10.1007/s10676-017-9430-8

[68] John Rawls. 1971. *A Theory of Justice*. Harvard University Press.

[69] Deborah L. Rhode. 1992. Ethics by the pervasive method. *J. Legal Educat.* 42, 1 (1992), 31–56.

[70] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 3 (Dec. 2015), 211–252.

[71] Khairulliza Ahmad Sallehab and Lech Janczewskia. 2016. Technological, organizational and environmental security and privacy issues of big data: A literature review. *Procedia Comput. Sci.* 100 (Jan. 2016), 19–28.

[72] Jeffrey S. Saltz, Neil I. Dewar, and Robert Heckman. 2018. Key concepts for a data science ethics curriculum. ACM Press, 952–957. DOI: https://doi.org/10.1145/3159450.3159483

[73] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. An algorithm audit. *Data and Discrimination: Collected Essays*. New America Open Technology Institute, New York, NY, 6–10.
[74] Paul M. Schwartz. 2011. Privacy, ethics, and analytics. *IEEE Secur. Priv. Mag.* 9, 3 (May 2011), 66–69. DOI: https://doi.org/10.1109/MSP.2011.61.
[75] Andrew D. Selbst and Solon Barocas. 2018. The intuitive appeal of explainable machines. *Fordham Law Rev.* 87 (2018), 1085–1138.
[76] Natasha Singer. 2018. Tech's Ethical'Dark Side': Harvard, Stanford, and Others Want to Address It. Retrieved from https://www.nytimes.com/2018/02/12/business/computer-science-ethics-courses.html.
[77] Michael Skirpan, Nathan Beard, Srinjita Bhaduri, Casey Fiesler, and Tom Yeh. 2018. Ethics education in context: A case study of novel ethics activities for the CS classroom. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education (SIGCSE'18)*. ACM, New York, NY, 940–945.
[78] Carol Spradling, Leen-Kiat Soh, and Charles Ansorge. 2008. Ethics training and decision-making: Do computer science programs need help? In *Proceedings of the 39th SIGCSE Technical Symposium on Computer Science Education (SIGCSE'08)*. ACM, New York, NY, 153–157. DOI: https://doi.org/10.1145/1352135.1352188
[79] Bernd Carsten Stahl, Job Timmermans, and Brent Daniel Mittelstadt. 2016. The ethics of computing: A survey of the computing-oriented literature. *Comput. Surveys* 48, 4 (Feb. 2016), 1–38. DOI: https://doi.org/10.1145/2871196
[80] Mark Staples, Liming Zhu, and John Grundy. 2016. Continuous validation for data analytics systems. In *Proceedings of the International Conference on Software Engineering (ICSE'16)*. ACM, New York, NY, 769–772. DOI: https://doi.org/10.1145/2889160.2889207
[81] Michael Steinmann, Sorin Adam Matei, and Jeff Collmann. 2016. *A Theoretical Framework for Ethical Reflection in Big Data Research*. Springer International Publishing, Cham, 11–27. DOI: https://doi.org/10.1007/978-3-319-28422-4_2
[82] Darren Stevenson. 2014. Locating discrimination in data-based systems. *Data and Discrimination: Collected Essays*. New America Open Technology Institute, New York, NY, 16–20.
[83] Julia Stoyanovich, Bill Howe, Serge Abiteboul, Gerome Miklau, Arnaud Sahuguet, and Gerhard Weikum. 2017. *Fides: Towards a Platform for Responsible Data Science*. ACM Press, 1–6. DOI: https://doi.org/10.1145/3085504.3085530
[84] John P. Sullins. 2017. *Ethics Boards for Research in Robotics and Artificial Intelligence: Is It Too Soon to Act?* Taylor Francis. DOI: https://doi.org/10.4324/9781315563084-5
[85] Christopher J. Sullivan and Michael G. Maxfield. 2003. Paradigmatic development in criminology and criminal justice: A content analysis of research methods syllabi in doctoral programs. *Crim. Just. Educat.* 14 (2003), 269–285.
[86] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Commun. ACM* 56, 5 (May 2013), 44. DOI: https://doi.org/10.1145/2447976.2447990
[87] Rong Tang and Watinee Sae-Lim. 2016. Data science programs in U.S. Higher education: An exploratory content analysis of program description, curriculum structure, and course focus. *Educat. Info.* 32, 3 (Jan. 2016), 269–290. DOI: https://doi.org/10.3233/EFI-160977
[88] Steven Tiell and Jacob Metcalf. 2016. *The Universal Principles of Data Science Ethics*. Technical Report. Accenture Labs.
[89] Rochelle E. Tractenberg, Andrew J. Russell, Gregory J. Morgan, Kevin T. FitzGerald, Jeff Collmann, Lee Vinsel, Michael Steinmann, and Lisa M. Dolling. 2015. Using ethical reasoning to amplify the reach and resonance of professional codes of conduct in training big data scientists. *Sci. Engineer. Ethics* 21, 6 (Dec. 2015), 1485–1507.
[90] Andrew Tutt. 2017. An FDA for algorithms. *Admin. Law Rev.* 69 (2017), 83–132.
[91] Shannon Vallor and Arvind Narayanan. [n.d.]. An Introduction to Software Engineering Ethics. Retrieved from https://www.scu.edu/ethics/focus-areas/more/engineering-ethics/an-introduction-to-software-engineering-ethics/.
[92] Casper Van Gheluwe and Sidharta Gautama. 2017. *Automated Data Quality Assessment for Citizen Science Platforms*. Master of Science in Computer Science Engineering. https://lib.ugent.be/catalog/rug01:002367458.
[93] Larisa Voronova and Nikolai Kazantsev. 2015. *The Ethics of Big Data: Analytical Survey*. IEEE, 57–63. DOI: https://doi.org/10.1109/CBI.2015.27
[94] Jun Jason Zhang Xinhu Zheng Xiao Wang Yong Yuan Xiaoxiao Dai Jie Zhang Wang, Fei-Yue, and Liuqing Yang. 2016. Where does AlphaGo go: From church-turing thesis to AlphaGo thesis and beyond. *IEEE/CAA J. Automat. Sinica* 3, 2 (Apr. 2016), 113–120.
[95] Richard Weiss, Michael Locasto, Jens Mache, Blair Taylor, Elizabeth Hawthorne, Justin Cappos, and Ambereen Siraj. 2015. Teaching security using hands-on exercises in 2015. In *Proceedings of the 46th ACM Technical Symposium on Computer Science Education (SIGCSE'15)*. ACM, New York, NY, 695–695.
[96] Joseph Weizenbaum. 1976. *Computer Power and Human Reason: From Judgment to Calculation*. W. H. Freeman.
[97] Janusz Wielki. 2015. The social and ethical challenges connected with the big data phenomenon. *Polish J. Manage. Studies* Vol. 11, No. 2 (2015), 192–202.

[98] Norbert Wiener. 1988. *The Human Use of Human Beings: Cybernetics and Society*. Perseus Books Group.
[99] Langdon Winner. 1980. Do artifacts have politics? *Daedalus* 109, 1 (1980), 121–136.
[100] Tal Zarsky. 2016. The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Sci. Technol. Hum. Values* 41, 1 (Jan. 2016), 118–132.
[101] Andrej Zwitter. 2014. Big data ethics. *Big Data Soc.* 1, 2 (July 2014), 205395171455925.

Received April 2018; revised October 2018; accepted December 2018