Intersectional AI: A Study of How Information Science Students Think about Ethics and Their Impact

NORA MCDONALD, University of Maryland, Baltimore County (UMBC), USA
SHIMEI PAN, University of Maryland, Baltimore County (UMBC), USA

Recent literature has demonstrated the limited and, in some instances, waning role of ethical training in computing classes in the US. The capacity for artificial intelligence (AI) to be inequitable or harmful is well documented, yet it's an issue that continues to lack apparent urgency or effective mitigation. The question we raise in this paper is how to prepare future generations to recognize and grapple with the ethical concerns of a range of issues plaguing AI, particularly when they are combined with surveillance technologies in ways that have grave implications for social participation and restriction—from risk assessment and bail assignment in criminal justice, to public benefits distribution and access to housing and other critical resources that enable security and success within society. The US is a mecca of information and computer science (IS and CS) learning for Asian students whose experiences as minorities renders them familiar with, and vulnerable to, the societal bias that feeds AI bias. Our goal was to better understand how students who are being educated to design AI systems think about these issues, and in particular, their sensitivity to intersectional considerations that heighten risk for vulnerable groups. In this paper we report on findings from qualitative interviews with 20 graduate students, 11 from an AI class and 9 from a Data Mining class. We find that students are not predisposed to think deeply about the implications of AI design for the privacy and well-being of others unless explicitly encouraged to do so. When they do, their thinking is focused through the lens of personal identity and experience, but their reflections tend to center on bias, an intrinsic feature of design, rather than on fairness, an outcome that requires them to imagine the consequences of AI. While they are, in fact, equipped to think about fairness when prompted by discussion and by design exercises that explicitly invite consideration of intersectionality and structural inequalities, many need help to do this empathy "work." Notably, the students who more frequently reflect on intersectional problems related to bias and fairness are also more likely to consider the connection between model attributes and bias, and the interaction with context. Our findings suggest that experience with identity-based vulnerability promotes more analytically complex thinking about AI, lending further support to the argument that identity-related ethics should be integrated into IS and CS curriculums, rather than positioned as a stand-alone course.

CCS Concepts: • **Social and professional topics** → Computing industry • **Computing methodologies** → Artificial intelligence

KEYWORDS

Artificial intelligence; education; ethics; algorithm bias; intersectionality

Author's addresses: Nora McDonald and Shimei Pan, University of Maryland, Baltimore County, 1000 Hilltop Circle Baltimore, Maryland 21250.

2573-0142/2020/October - 147 $15.00

https://doi.org/10.1145/3415218

ACM Reference format:

Nora McDonald and Shimei Pan. 2020. Intersectional AI: A Study of How Information Science Students Think about Ethics and Their Impact. *In Proceedings of the ACM on Human-Computer Interaction*, Vol. 3, No. CSCW2, Article 147, October 2020. ACM, New York, NY, USA. 17 pages. https://doi.org/10.1145/3415218

1 INTRODUCTION

Tech companies have long embraced the spirit of success in failure [2], best summed up by the motto "move fast and break things" [38]. University curriculums are, to some extent, working to reverse this thinking by challenging students to consider the ethical implications of AI-driven innovations like self-driving cars, and their personal roles in mitigating potential harms [33]. Recent security literature has highlighted the way in which data are the ultimate product, such that "data flows, algorithms and user profiling have become the bread and butter of software production" [14]. Scholars argue that, rather than adopt the EU General Data Protection (GDPR) laws, US privacy laws must account for power relationships between companies and culture which have normalized personal data collection and use for purposes that can be exploitative. [18]. Since power is a key element of this equation, the compounding of identities and vulnerabilities [4, 6, 7, 24] will inevitably shape the information relationships upon which we are dependent, and which put us at risk [26]. Of particular concern are the impacts of algorithms on vulnerable, intersectional identities—those whose race, class, nationality, gender or sexual identity, and other converging and compounding characteristics put them at particular risk in society at large. All these vulnerable constituencies share in common the fact that their identities make them more susceptible to further emotional, financial, physical harm, neglect, or discrimination by AI systems. The assortment of identities that flow from the concept of intersectionality is made exponentially larger by the number of dimensions that can create vulnerability and which, together, potentiate the impact of any one of them.

In fields like law and sociology, attention has been given to the known biases of algorithms affecting everything from our healthcare and social welfare benefits [3, 8, 23], to criminal justice [9, 27], to discrimination in the workplace [13], to Google search results [22]. We also know that these algorithm biases are inherited from our society [27]. While we expect that industry is grappling with the harms that algorithms impose on vulnerable populations, there is reason to be concerned about how we train incoming generations of developers for whom ethics education is, in some computer science departments, getting shorter shrift [36, 41]. Intersectionality in the context of AI development is an important focus of attention, not only because vulnerable groups require special protection to live safely, but also because keen sensitivity to extreme risk profiles better equips developers to think in imaginative ways about the protections needed by all.

This research is part of a larger project to explore AI ethics and implications for social justice and individual well-being in vulnerable intersectional communities. Through this work, we sought to understand what is happening at the university level to inform students about algorithm fairness; to learn whether students, themselves, worry about the impact of AI design on society and social justice; and to understand whether they consider particular implications for members of intersectional identity-based communities. Our research results indicate that while some of the students we studied are, in fact, attentive to the way that AI designs can be biased towards certain individuals and identities, they tend, on the whole, to believe responsibility for AI design bias and fairness implications resides with the company rather than with the designer. Students who are attuned to discrimination and social disadvantage are more sensitive to the role that context and structural inequality plays in producing bias, and also to the contributing role of model attributes and data training. The more attuned they are to structures of inequality, and the more "intersectionally-sensitive" they seem to be, the more complex and nuanced their thinking about AI bias and ultimate fairness. We also find that students believe the best way to mitigate bias is to have a human in the loop, usually an expert and/or someone with first-hand knowledge of the context. These findings, we argue, have implications both for curriculum design and student engagement in user research.

In the sections that follow, we situate our study in related literature, talk about our intersectional framing and concern for ethics in the workplace and classroom, discuss our findings from this qualitative study of information science graduate students, and assess the implications for identity-based learning and pedagogy. We conclude that our system of educating designers would profit from a closer integration of ethics education with technical training, as well as a sharper focus on vulnerable identities to produce more sensitive and proactive design thinking.

2 RELATED LITERATURE

2.1 Ethics and AI

Gillespie outlines six dimensions of relevance for ethical consideration in algorithm development: (1) "patterns of inclusion" (what data we put in); (2) "cycles of anticipation" (the relevance of the conceptualization of users and predictions of their behavior); (3) "evaluation of relevance" (the metrics and politics of decisions around including/excluding and weighing data and the implications for knowledge); (4) "promise of algorithmic objectivity" (the reliance/fall back to the concept of algorithms as being ideologically or politically neutral); (5) "entanglement with practice" (the way in which users' practices are shaped by the algorithms they use, sometimes in ways to subvert them); and (6) "production of calculated publics" (how algorithms' rendering of publics shapes the subjectivity of those publics and who is advanced by the knowledge) [11]. This analytic framework reminds us that the outcomes of algorithms are politically charged, driven by institutional decisions meant to perpetuate power structures.

Humans may be poor moral arbiters, not suitably wired to solve difficult ethical dilemmas, especially given the constraints of limited time and competition for resources [40]. Ethical considerations are inherently complex, and they are inevitably situated among other priorities. While it may be unreasonable to expect designers to fully *anticipate* the behaviors of code that have the potential to suppress freedom, it can be equally difficult for researchers to audit the empirical effects of algorithms after the fact. Furthermore, the very act of auditing may, itself, impose a normative view of what is just and appropriate behavior for an algorithm, creating challenges of a different sort from a different direction [29]. Who, then, is in a position to judge what is acceptable for any algorithm to do, and at what point in the process?

2.1.1 Bias and Fairness in AI. Discussion of AI bias and fairness is complicated by the fact that, though closely related, these two terms are still potentially distinct. The absence of intentional bias does not necessarily produce fairness, especially in an environment otherwise stacked against certain individuals; and conversely, designers of systems—whether socio-political systems or software systems—may find it necessary to employ certain kinds of bias to achieve fairness. Although beyond the scope of this paper, what is increasingly clear is that these terms cannot be contained within a "technical domain" [42] if they are to be applied to a discussion structural inequality. For the purposes of this paper, we consider bias to refer primarily to incorporation of stereotypes or prejudice into data collection or inputs which are the result of structural inequalities—through what they include and what they do not. (Limited sample size or limited information can also be sources of bias.) Unfairness is a characteristic of biased data inputs or systems, but it also describes the harmful or inequitable consequences of those biases when systems are deployed, favoring some at the expense of others. Removal of bias is a step toward fairness but a purely operational one that provides no guarantees as to actual outcomes.

By contrast, fairness, despite a multiplicity of definitions, grapples with outcomes: demographic parity, equal opportunity, and freedom from unwanted or discriminatory intrusions (e.g., [39]). A well-known issue in the field of AI is that being "fair" about an attribute, by removing it, can nonetheless yield highly discriminatory algorithms if there are other proxies in the database that point in the same direction. Algorithm application is also a critical dimension in assessing fairness. A "debiased" predictive algorithm that forecasts where crime will occur, if used by law enforcement to harass and oppress communities of color, becomes "unfair" by virtue of the use to which it is put [42]. Intersectionality, discussed later in this section, is a useful frame for thinking about AI fairness because it is specifically interested in the way that identities (or attributes) interact with inherited, structural bias, with the potential to produce unfairness, even where not intended.

2.2 AI and Vulnerable Populations

Discussion about AI fairness becomes even more provocative in the context of vulnerable identities—socio-economic characteristics, personal attributes, or life circumstances that put people at risk of discrimination or harassment. People with more privileged identities take for granted (whether or not they should) what they understand to be protective privacy norms; and they are often emboldened by those norms to relinquish identity information as a matter of course to the internet actors and private corporations seeking to monetize that information in various ways. The particular capacity of AI to do harm to vulnerable populations is well-documented [20, 43]. Their circumstances can make them more likely to encounter unsafe technologies and also to experience harmful consequences of varying sorts, including damage to their emotional well-being and subjectivity (how they believe others see and think about them) as well as risks to their economic and social well-being [19, 20]. When stereotypes about the welfare recipient are reinscribed in social welfare or healthcare systems, eligibility for social services can be adversely affected [3, 8]. These same stereotypes can also undermine privacy protection and produce stigma—for instance, by leading to requirements that poor women share information of doubtful relevance about their sexual history and personal relationships merely to receive social services [3]; or by re-encoding perceptions of Black and Latino young men into technologies of surveillance (e.g., gang databases) with direct implications for arrest and sentencing [9].

Sentencing algorithms are described as containing known racial biases such as "parental criminality," which reflect a long history of minority over-policing and racially charged arrests that are, in fact, "proxies for race" [43]. Richardson et al. discuss how predictive policing systems are built on data known to have been produced by biased (in some cases, unlawful) activities and policies [27]. These authors examined 13 jurisdictions and found evidence that the majority (nine out of 13) had been trained on so called "dirty data." In analytics and also algorithm development environments, that term commonly refers to data that are missing or wrong. These authors expand it to include "data that is derived from or influenced by corrupt, biased, and unlawful practices, including data that has been intentionally manipulated or 'juked,' as well as data that is distorted by individual and society biases" [27]. The consequences of dirty data and their capacity to mischaracterize vulnerable communities are such that we can expect from them the perpetuation, and perhaps exacerbation, of underlying inequities [8].

One pressing problem for vulnerable communities is the merging of existing data and facial recognition software, which has been demonstrated to be less accurate for minorities, particularly dark-skinned women [32] and children [12]. Despite Amazon's own acknowledgment that its facial recognition technology is not representative of the populations on which it is used [37], it continues to sell the product to government agencies (including law enforcement). Though not the only company to sell its facial recognition technology (e.g., IBM, Microsoft, etc.), Amazon has also teamed up with the US Immigration and Customs Enforcement (ICE) agency in a collaboration that, by integrating existing databases of public and private data, facilitates deportation [16].

2.3 Intersectional Lens

Any discussion of algorithm fairness must consider the concept of *intersectionality,* which argues for the importance of taking into account how race, class, gender, age, nationality, disability, and power structures (like capitalism, law, and policy) create added barriers to personal security and success when considered as multiplicative factors [39]. These identity characteristics can create compounding, complex experiences of inequality which are further influenced or exacerbated by structures of power and the social context [5]. Intersectionality demands that we grapple with the complexity of lives and circumstances and the dynamics of oppression [5].

The concept of intersectionality originates in the black feminist movement, and the phrase was "coined" [4, 5, 25] by Crenshaw in the late 1980s [6, 7], but it is gaining relevance today. Intersectionality is a powerfully evolving analytical framework, emerging now as an important element in critical social theory [4]. There are also numerous instances of its use in human-computer interaction (HCI) and computer science [35]. Though not explicitly, Eubanks evokes it in her exploration of how AI algorithms and prediction models, built on already racist policies, laws, and structures, may work to perpetuate discrimination through "algorithmic mutations" whose effects can be experienced by future generations [8]. Because intersectionality calls attention to the multiple dimensions through which fairness can be influenced and, thus, to the elevated odds of experiencing adverse effects, we find it a useful lens in studying how students think about this set of issues.

2.4 Critique of AI Developer Workplace Culture

A broad range of technologies use AI, many produced by large corporations whose objectives or priorities are aligned with their business goals. AI may be prone to bias because of the type of data it is fed or how the models are trained. While accountability is scarce, what few laws exist to confer protection do not directly deal with AI technologies [42]. Scholars have recently taken issue with development culture and its impact on AI development and bias. The AI Now Institute has identified central problems like this breach in accountability between those who design and use these technologies, and those who experience most harm as a consequence of them [42].

Scholars studying bias in AI have pointed out that these biases could be removed through relatively "modest" changes to formulas [1]. When groups have, for instance, different arrest rates based on racial bias, the result is a phenomenon called "predictive parity" whereby racial bias in outcomes is inevitable unless adjustments are made. Hardt et al. proposed such an adjustment, arguing that when algorithms merely ignore characteristics such as race, ethnicity or color, gender, etc., they act to encode and perpetuate biases that have origins elsewhere [17]. To avoid those malign effects, AI designers must not only consider their own biases, but those of society.

Revelations about the obliviousness of AI culture to the potential harms experienced by subgroups suggest a tone-deafness that may significantly reflect lack of workforce diversity, both in composition and perspective. The lack of institutional regulation to guide testing of algorithms for discriminatory or harmful effects further compounds the problem, leaving the AI organizational culture to its own devices [10]. In response to criticism, some companies have made a move toward diversity but concerns about tokenism and the authenticity or magnitude of change persist [41].

There are numerous high-profile cases of where AI went disastrously wrong because designers simply didn't imagine that their code would be used for nefarious purposes. In recounting the story of the tweet bot Tay.ai, Webb argues that the designers "relied only on their experience in China and their limited personal experience on social media networks. They didn't plan risk scenarios taking into account the broader ecosystem, and they didn't test in advance to see what might happen ..." [41]. Indeed, it would appear that developers may take a narrow, task-oriented view of their responsibilities which separates them from the very concept of responsibility for algorithm authorship. Notably, in discussions of "algorithms as culture" based on fieldwork conducted with a US developer of music recommendation algorithms, Seaver found that when asked what algorithm they worked on, company employees "located 'the algorithm' just outside the scope of their work, somewhere in the company's code" [31].

A variety of known strategies—adjusting algorithms to account for structural discrimination, taking into account/correcting for designers' own bias, and other socio-political dimensions of the environment in which they exist, and creating more diverse workplaces—are clearly essential to the development of fair algorithms and AI. In order to deploy them effectively, we must create an environment in which students are trained to reflect on these issues and ultimately, bring heightened ethical sensitivities into the workplace.

2.5 AI Ethics and the Classroom

Of Gillespie's six considerations, the ones with greatest relevance for students are the first four: the input data, conceptualizations of users and behavior, the politics of the decisions about what information to include or exclude, and the promise of impartiality. Yet, treatment of AI ethics in the classroom can be superficial or missing altogether [28]. In their survey of 186 ML-related courses at the top-20 computer science programs in the US, Saltz et al. found that only a little over one in ten ML courses include some ethics-related content, although they note that the number of stand-alone ethics courses offered in data science and AI programs is on the rise. A report from a National Science Foundation (NSF) funded project discussed the importance of incorporating ethics into core computer science curriculum, rather than as an elective ethics class [21].

3 STUDY DESIGN

This study was conceived to address the urgency and perils of a scenario in which future generations of software developers entering the workforce continue to bring to their jobs at major AI technology companies little or no ethics training to shape or restrain algorithm development. We approached this research with the following questions:

- How do students conceive of their own responsibility to the end-users (or targets) of AI system designs?
- How do students think about concepts of AI fairness and bias, and to what extent do they consider end-user identities and structures of inequality in that thought process?
- How do they think about ways to mitigate potential biases, and how do those lines of thinking relate to their own identity experiences and encounters with structures of inequality?

To address our research questions, we conducted semi-structured interviews with students recruited from one of two graduate-level courses on AI and Data Mining. Because Data Mining is often used to predict probabilistic outcomes from existing data, it is increasingly the locus of many ethical dilemmas.

Our discussion guide addressed: their studies and goals as graduate students in information science; their everyday experiences with AI and machine learning (ML); their concerns about privacy in AI; their concerns about impacts of privacy violations on the end-user; potential harms of AI technologies; concepts of AI fairness and bias; the importance of including people not like them in design; and perception of harms in a hypothetical design of a system used to predict whether students will succeed in their computer science or information science major. We started interviews by asking students about their encounters with AI in their everyday life, which often led them to talk spontaneously about bias related to those systems. They make the connection between bias and their identity as international students, as visa students, and non-native speakers and the conversation often flowed from there to discussions about AI bias and fairness and design. Each interview reflects the first researcher's conversations as it unfolded with participants, guided by student experience and familiarity with specific topics.

Our hypothetical design exercise involved asking students to imagine an AI design intended to predict student success in computer or information science majors. Students were asked to share their thoughts on:

- How such as system would work?
- What would be its main benefits and disadvantages?
- How might they overcome the disadvantages?

We have organized reporting of findings into three themes: (1) responsibility for bias and fairness, (2) mitigating bias, and (3) how students conceive of a hypothetical design. We plan to use these findings to inform future curriculum and research on this topic.

3.1 Interview Student Recruitment

Students were recruited from two graduate-level courses at a university to participate in a 60-90-minute interview over the phone or Skype (or whatever alternative medium they specified). Students were given information about the study on their course website and also through an in-class presentation. Students were offered the opportunity to either participate in the research or write a reflection essay for a single credit in their course. We provided a sign-up sheet on a

Google document but also gave them the option to sign up by individual email if they preferred. Table 1 details students by class and identifier used in quotations.

We did not take any identifying information from respondents, as is generally our practice to do everything possible to avoid or minimize identity consequences and potential harms to participants. While demographic information can be relevant in qualitative research analysis, here it is sufficient to know that, consistent with the makeup of the two classes, students who participated in the research were all of non-white ethnicity, primarily from Asian countries.

Table 1. students by class

	# of students	Interviews
AI class	11	[int 1-11]
Data Mining class	9	[int 12–20]
Total	20	[int 1-20]

Students in these classes had received no formal ethics training as part of this course curriculum, but we believe that participation prompted them to think about ethics in a way, and to a degree, that none of them appear to have done prior to the research.

3.2 Data Collection and Analysis

Interviews were recorded in all but two cases, where technical difficulties were encountered. Initial interviews were transcribed using Temi, a software that relies automated speech-to-text algorithms, but after a handful of interviews, this method was abandoned in favor of notes, since the transcription provided only conversational markers, not verbatim quotes, and was therefore deemed less useful. Available transcripts served to supplement memos and notes taken during and after the interviews.

Data were grouped into a hierarchy of themes relating to AI fairness. These themes are not representative of frequency, but rather, reflect a phenomenological approach whereby the researcher privileges participants' perceptions of phenomena [30, 34]. This phenomenological approach was taken partly in order to accommodate the language barrier that sometimes existed. Higher level themes were derived through comparative analysis and then thematic groupings were developed. The first author discussed initial themes with the second author to achieve further refinement. Results were memoed, and transcripts and notes were coded for initial themes.

4 FINDINGS

On their own, students do not tend to think deeply, or with great concern, about the implications of design for the privacy and well-being of others. Where they do, it is typically through the lens of their own identities, making that vantage point the easiest way to inspire expansive thinking about the implications of their designs.

Responsibility for removing or managing bias is something they tend leave with companies, who they believe have the financial power and political leverage to manage design applications on behalf of systems of power. When prompted to think about the implications, students tend to focus on bias, because it is an intrinsic feature of design, rather than on fairness, which is an outcome requiring them to imagine how and what consequences might actually occur. Their empathy can, however, be engaged in the consideration of fairness and they acknowledge that corporations won't necessarily take ownership of consequences, leaving no one on hand (except, some note, consumers themselves) to monitor outcomes. While the concept of fairness can be incorporated in their thinking, the notion of intersectionality is a more complex lens than they are currently inclined to use because it requires yet more empathy "work." They are better at projecting themselves than imagining the multifaceted complexity of others and the implications for heightened risk.

Students are aware of the existence of bias in AI design which disadvantages individuals based on race and gender (e.g., black women, immigrants) and they were able to articulate concerns about how these identities might make people more vulnerable to lost opportunities or other forms of unfairness. Since most of these students come from other countries, however, what is top of mind for them are biases that might influence opportunity loss for students on visas, non-native speakers, and immigrants.

In the sections that follow, we organize our findings around responsibility for the user, AI fairness, and mitigation of bias with consideration for identity-vulnerabilities and intersectionality and in so doing support the main thrust of our concluding argument that AI design learning must be integrated with ethics at every step.

4.1 Responsibility for AI Bias

Students think that "the company" designing the AI (as opposed to the designer or developers who work to operationalize it) is largely responsible for the user impacts and repercussions of AI, and they are aware of the myriad opportunities there are for such effects to occur. Students often talk about how Amazon Alexa's language processing limitations can influence search, loans, healthcare, and employee hiring algorithms, creating opportunities for active discrimination on a variety of groups that might be targeted for discrimination. There is no shortage of examples brought to the conversation.

One student points to the way companies like Facebook and Google use AI to manipulate users to keep using their systems and also the opportunities for discrimination those platforms or engines present. This student specifically references well-known example of racism in search, documented by Noble in *Algorithms of Oppression,* whereby a Google search for "unprofessional hairstyles for work" disproportionately returns pictures of black women [22]. The student associates this bias with stories connected to their own experience of being international, also citing a ride-sharing industry practice of using certain kinds of information to manage driver access to customers in a way that discriminates based on factors destined to lead to future opportunity loss.

Students provide other examples of the link between AI and international status in ways that can result in opportunity loss. Another recalls, while job-searching on a visa, receiving a lot of email and LinkedIn messages suggesting a highly targeted set of AI-driven outcomes that narrowed the job opportunities visible to them based on nationality and immigration status.

> "If there is a bias in the data, that will be reflected in the model as well ... While doing job search, I'm an international student right now, on a visa, while doing job search a lot of the information that is shared with me via email, or search results, or LinkedIn. I think they have some filters in which the company doesn't want to hire international students. For example, I can see that the number of replies that I get as international students, me or my friends, that is a lot less than maybe other residential students. I have a feeling that maybe there is a layer of AI that is filtering, which checks the

> resumes and maybe filters out. So that is the kind of hidden bias of the top of my head that I can think of … My options are extremely limited." [Int 11]

In this context, another student points out that more data is better: "if it's supervised learning, then we need all angles ... gender, race." [int 13].

The tendency to shift responsibility to end-users and their advocates was evident in a comment by one student, who speculates that, with increased transparency and awareness of the ways in which algorithms are biased, there will be opportunity for the market to react. This, they argue, will encourage or require companies to invest more time anticipating and solving problems proactively, even if the implications include slower time-to-market. On the same topic, another student points out that it is difficult to anticipate beforehand "what can go wrong" without empirical (post-market) evidence [int 13] but when a company that fails to take corrective action, it is directly responsible for its negligence.

Students occasionally argue that developers needed to assume some responsibility (" I think it's [also] the responsibility of the persons who are going to develop the system" [Int 4]). But even these students conceive of the developer as an agent without agency—partly because they do not have oversight or authority, and partly because they are not, themselves, biased and, therefore, prone to produce largely inadvertent effects which are overlooked, or even endorsed, by the power structures that direct their work.

A rather different take on the issue was expressed by respondents who feel that their own designs have integrity but who worry that a design made more cumbersome by additional protections might not be as widely used as a sleeker, more bias-prone, competitive product. In that scenario, measures intended to enhance protection vitiate the benefit of those protections. In the next section, where we discuss *fairness*, we see that students are more likely to take an active role in thinking about and embracing their own responsibility.

4.2 Responsibility for AI Fairness

While students seem to lay responsibility for algorithm bias at the feet of large corporations—particularly when those biases are linked to identity and structural inequalities—some nevertheless consider themselves capable of designing *fair* AI. When talking about their roles and responsibilities, students rarely use the word, bias (or even related operational terms like stereotyping, blind spots, etc.) because they appear to view the data as either "neutral" or outside their control, and the mathematical operations performed on the data to be neutral as well. Thus, when looking critically at AI, their focus is primarily on what use is made of data. This thinking is more aligned with concepts of fairness (e.g., equal opportunity, demographic and precision parity) than it is with a sensitivity to bias. Some do accede to the idea that the algorithms themselves may be problematic, but a lack of transparency makes it difficult to trace the root of the problem. One student touches on the way in which AI are unique to their maker, part of a black-box process that can result in misuse because its inner workings are not fully visible:

> "…so, organizations are grappling with explainability and transparency. It's all tied back to the machine learning life cycle, right? So, so I'm going to things that pop up in every company that I go into. How do I store my models? How do I map back to your original data that's associated with how I trained them? All right. Cause if you think about how many data scientists, machine learning experts are actually storing different

> versions of their models before they came up with their world-class model or their almost optimal model." [Int 5]

This student goes on to relate AI explainability to fairness, a complex social justice concept. A well-known issue in AI is that being "fair" about an attribute can yield highly discriminatory algorithms. Although fairness grapples with parity, equal opportunity, and accuracy, the implications for social justice require connection with the context and structural inequalities affecting individuals and communities. This student talks about using weather to model crime, observing that apparently non-biased AI can still have discriminatory outcomes.

Another student elaborates on the ways in which an algorithm can be unfair by recounting that Uber's AI manipulated workers into staying longer on their shift by giving them better rides after they attempted to sign off. Even though drivers may have made more money based on their willingness to work, this rose in the student's mind to the level of "bad AI" because "choice" [int 19] was being manipulated by an algorithm which, though not discriminatory per se and not even terribly complex, was manipulating driver behavior to the advantage of the company. In raising concerns, this student was pointing to policy and application more than to the way data were chosen or operationalized in an algorithm.

Some students take a literal view, considering AI to be equitable if assigns an equal distribution of outcomes, regardless of the implications.

> "So, if you divide by four, then twenty of my friends that benefit, like no bias. Like you give equal[ly] on each side." [Int 1]

There is, however, a general appreciation for the fact that what may be technically equal may not be truly equitable and that even literal equality of allocation is difficult to achieve because in complex models, some AI bias is inevitable. Several express ideas consistent with the view that "each human has bias, you cannot deny that" [int 12], leading some to suggest that mitigating bias requires diverse input to "screen" algorithms and multiple approaches with neutralizing effects. While students in the Data Mining Course were less likely to reflect on bias and fairness altogether, one did note that "machines are just trained as we train them" and thus, that we need to "train data on ethical and fair consequences" [int 19].

4.3 Mitigating Bias through Experts

Responsibility for bias has interesting fault lines, however. Some feel that progressing from equal allocation to true equity or fairness of outcomes may require the use of experts in a given area to oversee design. The concept of experts, as a tool of oversight, comes up repeatedly with students who are considering protocols for mitigating bias. One student considered that a panel designed to mitigate bias might include:

> "business, analysts, data analysts ... two to three people who can understand algorithms .... and one person with an arts background." [int 12]

On the other hand, students taking a more intersectional perspective (and who are also more attentive or sensitive to structural biases that people experience) are less interested in the role of expertise and theoretical review than real world empiricism. They are more likely to propose an interactive, outcome-based system whereby the algorithm would gain feedback in context, in the wild. While these students don't explicitly raise the issue of bias among experts, they take a more empirical approach, proposing to be guided by the actual behavior of AI than by predictions of potential behavior. This approach is perceived to circumvent or overcome human bias with robust real-world data that could be used to train the machine, although it does not necessarily deal with the paradox of bias in empirical outcomes assessment. At the same time, students with a particular sensitivity to social disadvantage as a contextual factor in dictating fairness outcomes are also led to consider the end-user as the relevant "human in the loop," offering an especially important source of insight on fairness outcomes.

> "Well, I can say that that the decision made by the human can be biased, but we have some way to correct our decisions, right? We can have feedback. So, if you are fully relying on the machine to make the decision it's not getting the feedback." [Int 2]

While not always explicitly looking to expertise as a solution, several others raise concern about the limitations of machine learning when faced with spurious correlations that limit opportunities for AI to be retrained in equitable or even effective ways without human intervention. A classic example, several students recount, is the correlation between ice cream consumption and drowning or, as one student offered, ice cream consumption and shark attacks, with potential to lead to incorrect causal inference. One observes the chicken and egg problem of training algorithms to use data effectively based on outcomes, noting the risk that those outcomes are, themselves, distorted by algorithm and data bias: "algorithm bias usually comes from incorrect parameter[s] or biased approach[es]" [int 9].

This student goes on to consider spurious correlations between cultural circumstances and behaviors, which work to perpetuate the outcomes that shape model input. That inherent circularity makes it difficult to identify the source of the bias. Despite an appreciation for the way in which data relationships can be spurious and misleading, most students characterize the algorithms that deploy them as essentially oblivious to the risks. AI doesn't understand right from wrong, they note, meaning that human review is essential to mitigate bias.

> "The data is not lying, right? The data is not necessarily telling you something that's wrong where the issues come in, and you don't have a full representation of the data that's out there. It was sampled wrong. The data might've been feature engineered wrong, where you might've did outlier detection and you remove some features that actually were important and now your model is a little off. So, number one, you can have the human review, the actual process for how the [algorithm] was, was created. All right, how did they do a feature engineer?" [Int 5]

We have discussed how students are thinking about responsibility and power, and their relationship to intersectional thinking. In the next section, we touch on some of the clear differentiators around bias as it relates to intersectional thinking.

4.4 Mitigating Bias through Intersectional Thinking

Students are prepared to consider the impact that social disadvantages have in producing AI bias, but those more attentive to intersectional barriers and structural inequalities were also the students more likely to consider the need for human interaction in mitigating bias, and to perceive the importance of user feedback to train the system:

> "But I think if we can make it more user specific, like the algorithm takes feedback from the user, this could be a better thing to do ...." [Int 2]

These students also more readily consider features and training sets as they relate to identity and discriminatory characteristics, revealing a link between their empathic and analytical reasoning as well as their willingness to assume responsibility. Several explicitly point out that the need for robustness in rule-building can be at odds with recognition of bias toward certain minority groups. One student thinks specifically of the way in which natural language processing (NLP) can be used to (unfairly) target minority groups like black activists, without taking into consideration their context. Interestingly, this student brought up concern for crime prediction algorithms, arguing that race sometimes gets used as a predictor of crime while overlooking income and access to opportunities (e.g., food, education, jobs).

These intersectionally-attuned students point out self-perpetuating biases associated with AI that uses historical datasets, with their patterns of bias and omission, to build models that guide future behaviors. They also observe that models are inherently error-prone from a statistical perspective (due to the irregularities and omissions of datasets, the challenges of feature-engineering, etc.) and that those errors may make models not just limited in their predictive accuracy but also simultaneously prone to unfairness. One student used the example of Michael Jordan, who—based on ordinary predictors—would not been singled out as a candidate for major league success because the model would not have been able to reflect key "intangibles."

> "A model would tell you that Michael Jordan was going to be horrible basketball player, right? If you would have asked Michael Jordan's father or some other individuals, [they'd say] I see his work ethic, his grit, his tenacity, his drive ... So, what you're really looking for is an assessment that can't be the intangibles that computers can't identify. And you're looking for the humans to be able to factor in those intangibles and cope." [int 5]

This kind of flexible and complex contextual thinking tends to go hand-in-hand with intersectional, analytical thinking among these students. The common thread is both an understanding of how algorithms are programmed to "see" the world through a privileged lens, and an appreciation for the unpredictability or inscrutability of the way models may behave—as well as the errors to which probabilistic modeling is inherently prone. It is for these reasons that sensitivity to cultural context and observation of algorithm behavior in the wild seem essential to mitigating bias.

4.5 Reactions to the Design Hypothetical

In designing a hypothetical algorithm intended to predict the success of students with information science or computer science graduate majors, students are sensitive to the challenges of evaluating students for whom English is not the first language, and they feel that such a model would, first, have to account for language-related biases like introversion. As a result of this sensitivity, they operationalize success in ways that are broader than class participation. They also attempt to create design structures that take into account biases that would be created if, say, changes in visa policies were overlooked. For instance, they argue that changes in policies that made it impossible for students to stay in the US after graduation would result in higher dropout rates and would thus bias the algorithm if unchecked.

> "I would consider ... how passionate they are and what kind of background they come from. Those two things would be my major concentrations." [Int 3]

It is only in the context of a hypothetical design exercise framed deliberately around their own experience that a broader set of students seem equipped to think about identity-related structural barriers (both obvious and subtle) and about intersectionality. For example, one student rehearses ways vulnerable elements of their own international identity and circumstances (language and visa status) might make them more susceptible to harm by third parties who might use those vulnerabilities to target and manipulate or exploit them.

> "Suppose they know I'm an international student. I'm easy to threaten. So, if a third party knows that I'm having trouble with finding a job, they may send more spams towards me then a native student. Also, they might threaten me based on my [status] that if, 'you don't do this you will lose your visa." [int 15]

There are a few exceptions to this sensitivity. A few point out that a feature like grades should take into account the area the student was from. In this way, they believe, structural biases can be mitigated.

> "And this is the problem, right? So, for a lot of these cases and use cases, the data for a lot of these individuals from these socioeconomically disadvantaged backgrounds, you have a lot of outliers right there. There are tons of outliers where the original information is not a representation of who they actually are." [Int 5]

While prepared to consider the input features for their algorithms, students are sometimes reluctant (or will neglect) to define "success" for the purpose of operationalizing the model outcome. Their tendency to think harder about the predictors than the predictions could be a function of limited training or a scholastic emphasis on feature engineering to predict already prespecified outcomes (with little or no agency to define them).

5 DISCUSSION

Students' capacity to think about AI bias is anchored in their personal experience of identity-based vulnerability. Alexa and Uber figure prominently in discussions about how AI can harm foreigners and, in particular, the way that our gig economy, when merged with AI, potentiates this harm. So, it is no surprise that students are quite familiar with how structures of discrimination in the US result in unfair treatment or mischaracterization of certain vulnerable groups. Some also express appreciation for the highly nuanced ways in which AI can be biased against themselves and against others. The tendency of people to draw heavily on their own experiences and perspectives in developing a broader worldview about bias and risk—a corollary of the broad principle that people generalize from their own experience—makes a strong argument for diversity among those charged with development. It suggests that encouraging diversity among developers will help produce a wider bulwark of protection against bias, and that workforce diversity claims should be looked at with a critical eye to ensure that they extend beyond mere tokenism.

In the following sections we discuss some implications for education in light of our findings, including some suggestions for the potential role of qualitative investigation, ethical discussions, and training that incorporates intersectional prompts into analytical thinking.

5.1 Implications of Responsibility for Users

While students believe "the company" has ultimate responsibility in AI bias, there are important implications to the finding that many already believe it is possible, potentially even desirable, to go beyond "expert" gatekeepers in pursuit of community feedback on algorithm performance and fairness. This suggests there may be some value in arming students with qualitative skills in order for them to gain crucial insight into those whom their algorithm has the potential to harm.

5.2 Mitigation and Intersectional Thinking

The more intersectionally-sensitive students were better able to connect identities with attributes and technical modeling considerations, suggesting that empathic reasoning and analytical thinking are complementary. Those with empathy or sightline into the identities and structural inequalities faced by others were more likely to think in a nuanced and analytical way about the features and attributes of their models, and also to consider the importance of collecting feedback from communities ultimately impacted by their systems post-market.

Even some of the students not readily inclined to approach the issues through this kind of frame could nonetheless envision the prospect that AI behavior might be misaligned with the developer's intention, simply by being prompted to access their own related experiences. The more students think about identities, or are inspired by conversation to do so, the more reflective they become about intersectional constraints. On their own, however, they may be less successful in utilizing their self-awareness to extrapolate empathetically beyond their own experiences to the circumstances of others.

The need to help many students relate familiar identity challenges to the complex intersectional vulnerabilities of others is a call to action for educators. Gillespie reminds us to think about the implications of algorithms in ways that can guide pedagogy toward a truly systematic approach to the integration of ethics with AI learning. The challenge for students in anticipating how algorithm behaviors may do inadvertent harm requires that students be made to contemplate algorithm properties and performance while simultaneously being encouraged to consider the identity vulnerabilities of others. That can occur only when ethics and algorithms are discussed in a coherent curriculum that demands ambitious thought experimentation from students.

5.3 Implications of Our Hypothetical

Students can think critically about proxies for vulnerable identities. What they cannot necessarily do is connect these concerns with the mechanics and outcomes of AI. The hypothetical serves to accentuate this disconnect between system and fairness. Notably, we had a student talk about explainable AI as the panacea, but they do not connect explainability with social justice. We see opportunity to expand explainable AI to go beyond merely supporting designers to helping the communities they affect [15].

This hypothetical showed the fragile nature of extrapolation. While we have already suggested that community engagement through qualitative study and access to students' own experience could be one approach, we also think that explainability could play an important role in helping students bridge bias and fairness.

6 CONCLUSIONS AND LIMITATIONS

Information Science graduate student display varying degrees of sensitivity to the way in which their identities interact with power structures to limit their opportunities. The more readily students make these connections, the more they worry about AI fairness, which in turn, makes them more interested in hearing from the people at risk than expert observers or intermediaries. This empathy "work" does not come easily. Students are willing to take on responsibility for the thought process around fairness, but they find it difficult to do and sometimes they need help.

We have provided justification and also guidance for a pedagogical approach to empathetic and responsible AI learning. This includes recommendations to prompt imaginative discussions, using an intersectional lens as well as suggested methods for extrapolation through engagement.

Our study is inspired by a sense of urgency about the way that AI are being used to shape every aspect of our lives with little consideration for the impacts to society. We are concerned about the hazards of an environment that apologizes later, and how those norms impact AI design.

We will use our findings to share future research about how best to prepare students to enter AI with a sense of empathy and concern for the most vulnerable, intersectional identities among us. Future research might study the impact of "live" case studies in the classroom with vulnerable individuals and its impact on the way that students design. We will be looking for ways to integrate intersectional thinking about AI into the curriculum and to study its impact on ethics throughout of students AI training.

Our work contributes novel insights about intersectional thinking in the context of AI bias and fairness. It provides guidance to scholars about how to build on these findings for pedagogy with integrative learning approaches. However, our study is limited by a small sample size of students who do not represent the full range of developers. Additionally, we encountered language constraints that may have limited the full articulation of their point of view and which also made it difficult for us to share our findings.

REFERENCES

[1] Angwin, J. and Larson, J. 2016. Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say. *ProPublica.*

[2] Appadurai, A. and Alexander, N. 2019. *Failure.* Polity.

[3] Bridges, K.M. 2017. *The Poverty of Privacy Rights.* Stanford Law Books.

[4] Collins, P.H. 2019. *Intersectionality as Critical Social Theory.* Duke University Press Books.

[5] Collins, P.H. and Bilge, S. 2016. *Intersectionality.* Polity.

[6] Crenshaw, K. 1989. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine. *University of Chicago Legal Forum.* 1989, 1 (1989).

[7] Crenshaw, K. 1991. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review.* 43, 6 (1991), 1241–1299. DOI:https://doi.org/10.2307/1229039.

[8] Eubanks, V. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin's Press.

[9] Ferguson, A.G. 2017. *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement.* NYU Press.

[10] Flick, C. 2016. Informed consent and the Facebook emotional manipulation study. *Research Ethics.* 12, 1 (Jan. 2016), 14–28.

[11] Gillespie, T. 2014. The Relevance of Algorithms. *Media Technologies: Essays on Communication, Materiality, and Society.* MIT Press.

[12] Goldstein, J. and Watkins, A. 2019. She Was Arrested at 14. Then Her Photo Went to a Facial Recognition Database. *The New York Times.*

[13] Guendelsberger, E. 2019. *On the Clock: What Low-Wage Work Did to Me and How It Drives America Insane.* Little, Brown and Company.

[14] Gurses, S. and Hoboken, J. van 2018. Privacy after the Agile Turn. *Cambridge Handbook of Consumer Privacy.*

[15] Hagendorff, T. 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines.* (Feb. 2020). DOI:https://doi.org/10.1007/s11023-020-09517-8.

[16] Hao, K. 2018. Amazon is the invisible backbone behind ICE's immigration crackdown. *MIT Technology Review.*

[17] Hardt, M., Price, E. and Srebro, N. 2016. Equality of Opportunity in Supervised Learning.

[18] Hartzog, W. and Richards, N.M. 2020. Privacy's Constitutional Moment and the Limits of Data Protection. *61 Boston College Law Review.* (Forthcoming 2020).

[19] Madden, M. 2017. *Privacy, Security, and Digital Inequality.* Data & Society.

[20] Madden, M., Gilman, M., Levy, K. and Marwick, A. 2017. Privacy, Poverty, and Big Data: A Matrix of Vulnerabilities for Poor Americans. *Washington University Law Review.* 95, 1 (Jan. 2017), 053–125.

[21] Martin, C.D., Huff, C., Gotterbarn, D. and Miller, K. 1996. Implementing a Tenth Strand in the CS Curriculum. *Commun. ACM.* 39, 12 (Dec. 1996), 75–84. DOI:https://doi.org/10.1145/240483.240499.

[22] Noble, S.U. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism.* NYU Press.

[23] O'Neil, C. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Broadway Books.

[24] Pierce, J., Fox, S., Merrill, N. and Wong, R. 2018. Differential Vulnerabilities and a Diversity of Tactics: What Toolkits Teach Us About Cybersecurity. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018), 139:1–139:24. DOI:https://doi.org/10.1145/3274408.

[25] Rankin, Y.A. and Thomas, J.O. 2019. Straighten Up and Fly Right: Rethinking Intersectionality in HCI Research. *Interactions.* 26, 6 (Oct. 2019), 64–68.

[26] Richards, N.M. and Hartzog, W. 2017. Privacy's Trust Gap. *126 Yale Law Journal.* 1180, (2017).

[27] Richardson, R., Schultz, J. and Crawford, K. 2019. Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. *New York University Law Review Online.* (2019).

[28] Saltz, J., Skirpan, M., Fiesler, C., Gorelick, M., Yeh, T., Heckman, R., Dewar, N. and Beard, N. 2019. Integrating Ethics Within Machine Learning Courses. *ACM Trans. Comput. Educ.* 19, 4 (Aug. 2019), 32:1–32:26. DOI:https://doi.org/10.1145/3341164.

[29] Sanvig, C., Hamilton, K., Karahalios, K. and Langbort, C. 2014. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry* (Seattle, Washington, USA, May 2014).

[30] Schutz, A. 1967. *The Phenomenology of the Social World.* Northwestern University Press.

[31] Seaver, N. 2017. Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society.* 4, 2 (Dec. 2017), 2053951717738104. DOI:https://doi.org/10.1177/2053951717738104.

[32] Singer, N. 2019. Amazon Is Pushing Facial Technology That a Study Says Could Be Biased. *The New York Times.*

[33] Singer, N. 2018. Tech's Ethical 'Dark Side': Harvard, Stanford and Others Want to Address It. *The New York Times.*

[34] Starks, H. and Brown Trinidad, S. 2007. Choose Your Method: A Comparison of Phenomenology, Discourse Analysis, and Grounded Theory. *Qualitative Health Research.* 17, 10 (Dec. 2007), 1372–1380. DOI:https://doi.org/10.1177/1049732307307031.

[35] Thomas, J.O., Joseph, N., Williams, A., Crum, C. and Burge, J. 2018. Speaking Truth to Power: Exploring the Intersectional Experiences of Black Women in Computing. *2018 Research on Equity and Sustained Participation in Engineering, Computing, and Technology (RESPECT)* (Feb. 2018), 1–8.

[36] Thompson, C. 2019. *Coders: Who They Are, What They Think and How They Are Changing Our World.* Picador.

[37] Thoughts On Machine Learning Accuracy: 2018. *https://aws.amazon.com/blogs/aws/thoughts-on-machine-learning-accuracy/.* Accessed: 2019-07-31.

[38] Velazco, C. 2018. Facebook can't move fast to fix the things it broke. *Engadget.* (2018).

[39] Verma, S. and Rubin, J. 2018. Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness* (Gothenburg, Sweden, May 2018), 1–7.

[40] Waser, M.R. 2015. Designing, Implementing and Enforcing a Coherent System of Laws, Ethics and Morals for Intelligent Machines (Including Humans). *Procedia Computer Science.* 71, (Jan. 2015), 106–111. DOI:https://doi.org/10.1016/j.procs.2015.12.213.

[41] Webb, A. 2019. *The Big Nine: How the Tech Titans and Their Thinking Machines Could Warp Humanity.* PublicAffairs.

[42] Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S.M., Richarson, R., Schultz, J. and Schwartz, O. 2018. *AI Now 2018 Report.* AI Now Institute.

[43] 2018. *Litigating Algorithms: Challenging Government Use of Algorithmic Decision Systems.* AI Now Institute.

Received January 2020; revised June 2020; accepted July 2020