The Future of AI Can Be Kind:
Strategies for Embedded Ethics in AI Education

Yim Register

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:

Emma S. Spiro, Chair

Jevin West

Wanda Pratt

Program Authorized to Offer Degree:

The Information School

University of Washington

Abstract

The Future of AI Can Be Kind: Strategies for Embedded Ethics in AI Education

Yim Register

Chair of the Supervisory Committee:

Emma S. Spiro

The Information School

The field of Data Science has seen rapid growth over the past two decades, with a high demand for people with skills in data analytics, programming, statistics, and ability to visualize, predict from, and otherwise make sense of data. Alongside the rise of various artificial intelligence (AI) and machine learning (ML) applications, we have also witnessed egregious algorithmic biases and harms – from discriminatory outputs of models to reinforcing normative ideals about beauty, gender, race, class, etc. These harms range from high profile cases such as the racial bias embedded in the COMPAS recidivism algorithm, to more insidious cases of algorithmic harm that compound over time with re-traumatizing effects (such as the mental health impacts of recommender systems, social media content organization and the struggle for visibility, and discriminatory content moderation of marginalized individuals [400, 401]). There are various strategies to combat and repair algorithmic harms, ranging from algorithmic audits and fairness metrics to AI Ethics Standards put forth by major institutions and tech companies. However, there is evidence to suggest that current Data Science curricula do not adequately prepare future practitioners to effectively respond to issues of algorithmic harm, *especially* the day-to-day issues that practitioners are likely to face. Through a review of AI Ethics standards and

the literature, I devise a set of 9 characterizations of effective AI ethics education: *specific, prescriptivist, action-centered, relatable, empathetic, contextual, expansive, preventative, and integrated*. The empirical work of this dissertation reveals the value of embedding ethical critique into technical machine learning instruction – demonstrating how teaching AI concepts using cases of algorithmic harm can boost both technical comprehension and ethical considerations [397, 398]. I demonstrate the value of relying on real-world cases and experiences that students already have (such as with hiring/admissions decisions, social media algorithms, or generative AI tools) to boost their learning of both technical and social impact topics. I explore this relationship between personal relatability and experiential learning, demonstrating how to harness students' lived experiences to relate to cases of algorithmic harm and opportunities for repair. My preliminary work also reveals significant *in-group favoritism*, suggesting students find AI errors more urgent when they personally relate to them. While this may prove beneficial for engaging underrepresented students in the classroom, it must be paired with empathy-building techniques for students who relate less to cases of algorithmic harm, as well as trauma-informed pedagogical practice. My results also revealed an over-reliance on "life-or-death reasoning" when it came to ethical decision-making, along with organizational and financial pressures that might impede AI professionals from delaying harmful software. This dissertation contributes several strategies to effectively prepare Data Scientists to consider both technical and social aspects of their work, along with empirical results suggesting the benefits of embedded ethics throughout all areas of AI education.

Acknowledgments

This work would not have been possible without my various mentors throughout this Ph.D. program. I am so grateful to Emma S. Spiro, who has supported every one of my bizarre ideas with genuine curiosity – your support and brilliance has helped shape me into a better scholar and scientist. Meeting with you *always* set me on a better path, and your commitment to your students and our various identities and interests does not go unnoticed. My Ph.D. experience would not have been the same without Jevin West; I could not ask for a better co-teacher and mentor – it has been my joy to laugh so hard with you in a machine learning classroom. I was able to bloom into the academic I always wanted to be because of the experiences in our Data Science II course. Thank you so much to Wanda Pratt, who guided me when I was most lost – shaping my interests towards advocacy and user agency, as well as believing in me as a young trans academic just trying to make sense of science. Thank you Amy Zhang, my GSR, for your genuine interest in my work and thoughtful questions and support. Special thanks to Amy J. Ko for welcoming me into the world of computing education research, and laying the foundation for me to thrive. My Ph.D. has also been punctuated with delightfully creative internships and side quests, a sincere and joyful thank you to Dan Schneider and Greg Wilson: who showed me that work can be *fun*! Deepest gratitude to Nic Weber, who has quickly become a guiding light helping me across the finish line. Thank you to Anna Lauren Hoffman, who showed me that pursuing technology justice is one way I can bring love into the world. Many thanks to my friend Julia Dunbar for being by my side throughout this whole process – the rejections and difficulties were all made better because you were there. Thank you to Yolanda Barton, for showing me that even our wildest dreams are possible when they come from the heart. Thank you to Xiaobing Xu, Nayan Kaushal, Joseph William Tan Garcia, and Dev Wilder, for surviving my first ever Directed Research Group – and for inspiring me in more ways than I can describe. Thank you to my entire cohort and fellow PhD students, who have laughed with me and debated the very nature of existence, only to shrug and hope for the best even if reality is meaningless.

I am beyond grateful to the National Science Foundation for funding my graduate education, as well as the UW iSchool for providing teaching and research opportunities along the way. Thank you to the Center for an Informed Public for welcoming me and including me; especially to Liz Crouse and everyone involved in Misinfo Day!

Finally, to my students. Each and every one of you taught me how to be a better teacher. Whether through successes or failures, thank you for giving me the opportunity to learn how to be better. It has been my joy to see you grasp difficult concepts, advocate for yourselves, empower others, and speak up about social injustices. Keep going, you belong here. It will be because of you that the future of AI can be kind.

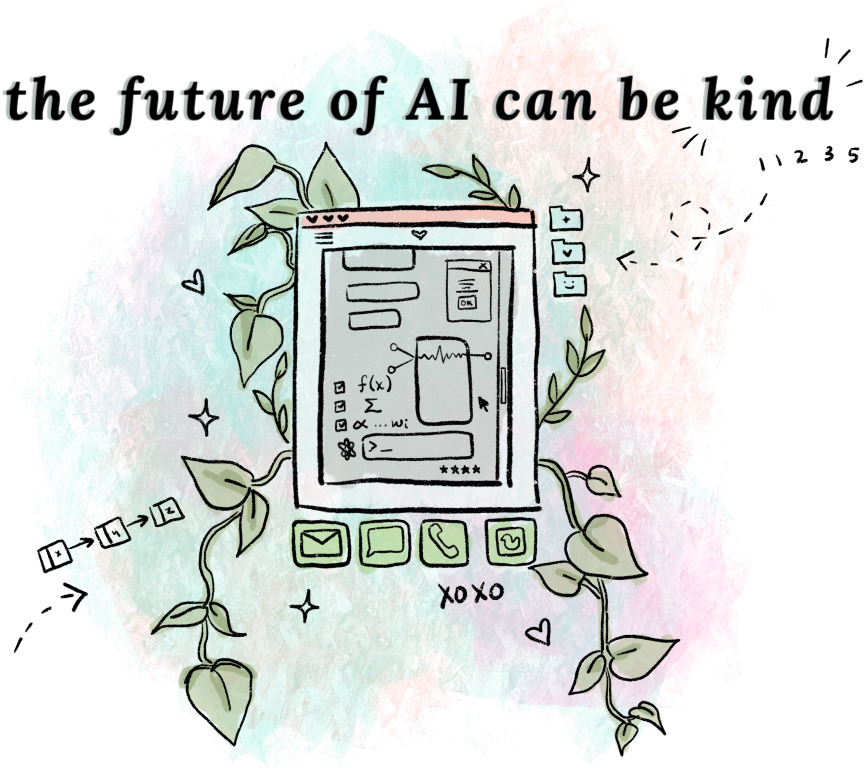
Dedication

To every single person who has loved me along the way.

And to Nick Duran, who taught me how to love myself.

The Future of AI Can Be Kind:

Strategies for Embedded Ethics in AI Education



Contents

Part 0
Introduction

"How do we hold people accountable for wrongdoing and yet at the same time remain in touch with their humanity enough to believe in their capacity to be transformed?"

– bell hooks

Over the past decade, Data Science has grown rapidly, with more and more demand for data professionals [324, 244]. Data Science includes machine learning (ML) and artificial intelligence (AI) skills, which have been applied in nearly every sector *(i.e. medicine, finance, education, manufacturing, commerce, judicial systems, entertainment, etc.)*. We exist now in a generative AI boom, and the demand for AI skills continues to increase. To prepare future Data Scientists and AI professionals, universities now offer a myriad of Data Science programs, from undergraduate to PhD levels. In Chapter 1, I review current curricula and priorities for these Data Science programs. One area of Data Science curricula that needs improvement is preparing our students to handle algorithmic bias and algorithmic harm [357], particularly as we witness more and more cases of AI mistakes. For example, in 2016 the audit of recidivism risk scoring algorithm COMPAS demonstrated racial bias, consistently labeling Black offenders as 'higher risk' than their white counterparts; even if their crimes were similar [21]. The 2018 paper *Gender Shades* revealed disproportionate inaccuracies of facial recognition software on Black female faces [80]. This brings us to today, where generative AI also replicates biases, such as in image generation perpetuating racial and gender stereotypes, especially for different occupations like *engineer* or *janitor* [351]. In many cases, the AI mistakes or biases are subtle and insidious, for example, the promotion of pro-anorexia content through social media recommender systems [465] or discriminatory content moderation that disproportionately bans LGBTQ speech [365, 212, 401]. In Chapter 2 I provide a casebook (See Table 5) of algorithmic harm examples across industries and technologies.

We can agree that many of these harms are not the systems working as intended; and need to be repaired. Either because of moral obligation, values, customer satisfaction, or simply to avoid litigation and comply with regulation – we want our Data Scientists to be able to respond to issues of algorithmic harm. It may even be as simple as raising awareness of the possibilities of such harms, and recommending auditing practices to avoid unwanted outcomes [396].

One resource already in place are Codes of Ethics – ethical standards for AI that guide how we create our AI technology in responsible ways. The ACM, the United Nations, the White House, and Big Tech companies like Google/Microsoft/Amazon/Meta/IBM have all put forth AI ethics standards. I review the commonalities between these standards in Chapter 3. These ethical standards tend to list normative recommendations like "avoid harm", "be fair", "be trustworthy", and "contribute to society". Unfortunately, there is evidence to suggest that these kinds of ethical guidelines don't translate into practice and are difficult to operationalize [332, 333, 209, 172]. For example, one study compared a group of developers using the ACM Code of Ethics with a control group, and there was *no difference* in the software decisions they made [314].

You can imagine why this might be. Consider the recommendation of "avoid harm". It would be much more meaningful to provide an example of algorithmic harm, such as bias in a skin cancer detection model [204], and demonstrate how to look at the confusion matrix of false positives and false negatives for different demographics in your data. You might then also demonstrate how to tweak a model to prioritize recall over precision, or the importance

of balancing the training data. These decisions may improve outcomes for the vulnerable population, while simultaneously covering technical details of careful and effective Data Science. You can imagine more practical examples like this for "be fair", or "protect privacy" or "be socially responsible", and actually give students the tools and hands-on examples to combat algorithmic harms, as well as deeply investigate the details of their models. Many scholars have put forth recommendations for effectively teaching AI ethics, including Casey Fiesler, Deborah Raji, Amy J. Ko, and Emily Bender, just to name a few. In Chapter 3 I summarize the literature on operationalizing AI ethics into 9 characteristics of successful AI ethics education: It is *specific, prescriptivist, action-centered, relatable, empathetic, contextual, expansive, preventative,* and *integrated*.

These characteristics boil down to three pertinent ideas: 1) AI ethics ought to rely on real world cases already rich with detail, 2) those cases should appeal to student's lived experiences with day-to-day algorithmic systems, and 3) ethics should not be left to the end, or 'othered' as a soft skill, and should instead be *integrated* throughout the technical curricula.

Presumably, if we follow these normative claims, something should be "better" in some way. But how do we define "better" in this case? Will it improve learning outcomes? Will we increase representation of marginalized groups in AI? Will we see more empathy? What will we get out of integrating AI ethics into technical AI coursework that we didn't have before?

The research presented in this dissertation provides empirical evidence to support the normative assertions about AI ethics. What happens when we use student's lived experiences to teach about AI? What kinds of algorithmic harms should we explore, and what does that help in terms of learning outcomes? What do students need in order to support their ethical decision-making in practice?

First of all, "lived experiences" with day-to-day algorithms looks different for all of us. However, we all share some common ground, as AI exists right in our pockets and laptops. However, the AI instruction that I received in my undergraduate studies looked more like abstract equations and toy examples, despite the wealth of examples we could be relying on to teach AI in relatable ways. For example, we have all interacted with recommender systems on social media or Google Scholar. And we have been tracked or quantified by a variety of systems from our Apple Watch to our college applications. According to best practices in computing education, we can rely on lived experience and experimentation to better understand the systems we want to learn about. This is proposed by both *constructionism* [377, 374] and *experiential learning*[139, 286], two theories that emphasize the importance of self-driven projects and direct experiences with the subject matter at hand.

I wanted to know if including such lived experiences or situated perspectives of AI did anything for learning the technicalities of AI systems, as well as to bring awareness to potential downsides of the AI systems we create. 1) could it help students actually learn the technical pieces as constructionism suggests? and 2) could it make them more attuned to possible cases of algorithms gone wrong, preparing them to think about AI safety concerns?

In my first study presented in Chapter 1, I test the effect of using personal data and lived experience to learn a technical concept: linear regression with gradient descent. We measured both comprehension of the technical mechanisms *and* how students applied these in advocacy arguments in the case of a faulty financial aid algorithm. We found that people who used personal data advocated with more technical mechanisms, more attention, and more cohesive advocacy arguments. Therefore, it may be the case that using personal data and lived experiences to teach AI provides a boost in technical understanding.

In a follow-up study presented in Chapter 3, I taught collaborative filtering within the context of social media ad recommendation, broadening our participant pool to social media users in general (many of whom were also Data Science students, which we compared to non-expert users). Users looked at their own Facebook ad data in a web app designed to show them

how recommender systems work. Similar to the previous study, I wanted to know if using one's own personal data to learn about the mechanisms of an algorithm would impact how well a participant surfaced issues of algorithmic harm. Are they more attuned to these issues when using their own data? And can this be done while simultaneously teaching technical mechanisms? We found that this was a successful way to both teach the technical mechanisms of recommender systems and to surface various problems on social media like: pro-anorexia content, political polarization, misinformation spread, LGBTQ privacy risks, the inability to appeal content moderation decisions, and the impact of 'hate following' a topic. Participants who successfully understood the technical mechanisms were more likely to use them to explain how algorithmic harms occurred, and use them to propose potential solutions.

We also saw evidence of lived experience as a booster to technical knowledge in non-experts trying to figure out the Instagram algorithm. In my work regarding algorithmic precarity, we saw sophisticated analysis of the Instagram algorithms by creators who had lived experiences of content moderation, privacy breaches, targeted advertising, or unpredictable engagement metrics Register et al. [400]. For example, users would explain why the algorithm favored some things over others, how bot accounts impacted the likelihood of engagement, or why certain images or speech patterns get banned over others. This directly relates to the algorithmic sensemaking literature, which we cover in Chapter 2.

In these various research studies, we demonstrate that lived experience seems to give a boost for learning AI that can help learners with both technical and ethical details. However, with the above studies I only looked at scenarios that were universally relatable for all my participants: university metrics for university students, social media recommendations for social media users, content moderation for content creators. While we saw the benefits of including lived experiences in AI education, what if students do *not* have lived experience of a problem? This may be especially pronounced when the demographic makeup of the student body is more homogeneous and lacking diversity, as suggested by the AI Index Report showing gender and racial gaps in the demographic makeup of the AI sector [530]. By personalizing learning to everyone's own point of view, will we only ever be interested in things that personally affect us? Some scholars refer to this as *"in-group favoritism"* and suggest this as a main barrier to mitigating algorithmic harms [209]. In-group favoritism suggests that many issues of algorithmic harm get overlooked when they don't directly impact the dominant group.

This is a testable notion. Do our students exhibit in-group favoritism when it comes to cases of algorithmic harm? While there is some rhetoric that Data Scientists "don't care about ethics" we are seeing more evidence that this is not entirely true, though curricula have historically prioritized other topics [169]. I would be surprised if we asked students about cases of algorithmic harm and they said they did not care at all – but it may be the case that some students care more than others. Who are they, and why?

In Chapter 4 I present the results of a preliminary study that investigates in-group favoritism and ethical decision-making for 30 different cases of algorithmic harm, many of which are drawn from the casebook presented in Table 5. Students rated different scenarios for how relatable they are, how urgent they are, whether or not they would commit to working on such problems, and whether or not the software should be deployed in an industry scenario. We found evidence of in-group favoritism, as well as some evidence for *the bystander effect*: students might find something urgent (but not relatable), and therefore assume someone else would be a better fit to work on it. We also found that decision-making was most influenced by extremes, students primarily analyzed the severity of harm by whether or not it was a clear life-or-death scenario.

At the end of the day, Data Scientists make decisions and must justify those decisions. While of course our students should be prepared to audit software that dictates whether or not someone is arrested or given cancer treatment, most Data Science students will not be tasked with life-or-death decisions so glaringly egregious. Instead, they will face more nuanced,

subtle, and insidious mistakes with downstream effects and complex tradeoffs. For example, my content moderation work points to the downstream effects of automated content moderation which produces discriminatory outcomes (presented in Chapter 2. Our students tend to go off into Data Science roles of e-commerce, social media companies, and more and more of them will be headed to work on generative AI product teams. Issues of bias in these settings may go easily undetected, and require a technical eye to really monitor the inputs, outputs, performance, and tradeoffs of a particular AI system. So, let us actually prepare them with the skills, tools, and confidence to do those jobs well *and* keep ethics in mind. Our students are uniquely situated with the technical details to be able to notice issues in their algorithms and hypothesize why their models produce errors. It gives us a chance at a more thoughtful and kind future if we give practitioners things they can actually use to ensure AI safety: real-world case studies, relatable examples, auditing techniques, evaluation schemas, and project-based exploration of complex problems. Drawing from literature on responsible AI, value-sensitive design, STS, computing education, and learning sciences I have contributed different strategies for embedding ethics into technical AI instruction.

I see this as a way to refocus us to the good that our technology can do. When I co-taught Master's level machine learning students through the lens of social impact and ethical consideration, they produced thoughtful final projects that linked algorithmic knowledge with data science for good. For example, in creating a model for determining walkability scores, a student wrote *"these models dont generalize to everyone, and there could be poor representation for children, disabled people, and the elderly. We need to adjust the model to be more inclusive and specifically serve their needs"*. Another student did a project on heart disease prediction, and wrote *"there is a long standing history of medical studies only collecting health data from white males. This can be highly consequential for diagnosis and treatment. It is vital these datasets are more representative."* These reflections accompanied self-driven projects on highly technical material, which historically did not ask for such ethical consideration at all. Other self-directed projects included stroke prevention, fentanyl detection, content moderation fairness, and suicide risk prediction. This particular *Choose Your Own Machine Learning Adventure* lesson was highlighted in Register et al. [399].

Educators may be daunted by the idea of embedding ethics into technical instruction, as many cases of algorithmic harm involve race and gender-based discrimination. Through this work I have developed the LEADERS Framework to guide educators to teach about AI societal impact. This framework has been adopted in a professional development course from Code.org. I provide the details on the LEADERS Framework in 5 along with several other educational resources and insights. In summary, the framework encourages educators to rely on real-world cases, always assume your datapoint is in the room, and take on a *solution-oriented* approach that helps students feel empowered both technically and ethically.

Future work will be to empirically test the efficacy of different strategies that impact ethical decision-making in a Data Science role. In Chapter 4 I discovered that students overly relied on life-or-death as their way to make decisions, and only felt confident when there was clear bodily harm or egregious unfairness. However, many cases of algorithmic harm are far more subtle. By exploring the barriers to ethical decision-making and to look for common pitfalls, we may be able to develop more robust guidelines and educational strategies for preparing future Data Scientists to carefully investigate the models they create. Not only will this be good for society, but it is simply *good Data Science*.

✓ **Summary**

The goal of this dissertation is to provide evidence that the use of situated learning approaches for AI education can also help us teach ethics and advocacy alongside technical concepts at advanced levels. Not only do I argue the efficacy of situated learning but I contribute several ways to center it in AI instruction, including empirical results on how relatability to cases of algorithmic harm spurs interest and care. I finish with an outline of a framework for teaching ethics in an embedded way, with potential lessons motivated by my empirical findings.

I contribute 4 verbatim peer-reviewed research studies as well as the results from a preliminary study currently in preparation.

- Yim Register and Amy J. Ko. "Learning Machine Learning with Personal Data Helps Stakeholders Ground Advocacy Arguments in Model Mechanics". In: *Proceedings of the 2020 ACM Conference on International Computing Education Research*. ICER 20. Virtual Event, New Zealand: Association for Computing Machinery, 2020, pp. 67–78
- Yim Register et al. "Attached to The Algorithm: Making Sense of Algorithmic Precarity on Instagram". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–15
- Yim Register et al. "Beyond Initial Removal: Lasting Impacts of Discriminatory Content Moderation to Marginalized Creators on Instagram". In: *Computer Supported Cooperative Work (CSCW)* 8 (2024)
- Yim Register and Emma S Spiro. "Developing Self-Advocacy Skills through Machine Learning Education: The Case of Ad Recommendation on Facebook". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 16. 2022, pp. 817–828

Part 1
Data Science: Industry Demand And Education Opportunities

"Our own values and desires influence our choices, from the data we choose to collect to the questions we ask. Models are opinions embedded in mathematics."

– Cathy O'Neil, *Weapons of Math Destruction*

Abstract

Data Science and AI continue to grow, with high demand for professionals who can make sense of the vast amounts of data available – however we do not yet know best practices for teaching Data Science, especially in ways that adequately prepare practitioners to combat a wide variety of algorithmic harms. One solution is to integrate what we know from computing education research, and rely on situated knowledge to help students better connect with their data and the outputs of the algorithms they may create.

With technological advancement over the past three decades, we have seen an enormous increase in the amount of *data* recorded and processed. In order to make sense of and derive value from such data, we need data professionals who can select, analyze, and make informed decisions from the copious amounts of data available. The early 2010s witnessed numerous calls for more *Data Scientists* [304, 127, 371] – people who could make sense of data for predictive analytics, decision-making, and for the underlying algorithms that power technologies across nearly every sector: social, economic, medical, educational, judicial, creative, manufacturing, etc. (*data is everywhere!*). In response, we saw a boom in Data Science programs for higher education [229, 216, 55, 17], and in recent years more efforts towards AI education for all ages. This chapter summarises the core components of current Data Science curricula in higher education, demonstrating an opportunity for a more proactive way of integrating ethics and responsibility in how we teach Data Science and AI. This chapter argues from curricula review, as well as position papers, that most ethics components are either left out or 'othered' in some fashion [395, 180]. In order to approach teaching ethics in context, we draw upon best practices in computing education: introducing educational theories such as experiential learning [139], constructionism [374, 375], and situated learning [286] in computing contexts and beyond. This chapter demonstrates how to employ those learning theories in practice where learners use their *own data* in context of an AI algorithm lesson. The results suggest that this method improves comprehension of the technical content and engagement in a self-advocacy task [397]. These findings lay the foundation for teaching technical skills in parallel with ethical concepts, using a situated learning approach, which we build upon in the following chapter.

1.1 Author's Preface

I was born at the start of the world-wide-web. Google hadn't been invented yet, and we were still on dial-up. "Digital connection" meant a series of hyperlinks between HTML pages. I learned to crawl around the same time that PageRank did. I remember MySpace, I remember AOL Instant Messenger, I remember incessantly Googling how to cheatcode my way to some serious cash for my Sims. I remember being warned against using Wikipedia, and should turn to 'primary sources' instead. I remember how quickly it became a joke that we shared whatever thoughts popped into our minds in our online status updates. I remember posting anyways.

I grew up as the Internet grew up.

1997 was the year that IBM's DeepBlue beat then reigning chess champion Garry Kasparov in a series of matches that shocked the world; I was three. Google launched to the public in 1998, 26 years ago. Facebook launched in 2004, followed by YouTube (2005), Reddit (2005), Twitter (2006), and later Instagram (2010), Pinterest (2010), Snapchat (2011), and TikTok (2016). The Netflix Prize ran in 2009 – a competition for producing the best collaborative filtering (recommender) algorithm. Facebook began using facial recognition for tagging images in 2010. IBM Watson went up against Jeopardy champion Ken Jennings in 2011. Tesla announced the Autopilot feature for the Model S in 2014, the same year that ALEXA was released. Continuing our collective fascination with games, the next challenge was *Go*; Google's AlphaGo successfully beat world champion Lee Sedol in 2016. In 2017, the FDA approved an AI-assistive technology to help radiologists better spot breast cancer. Picking up in 2018, Natural Language Processing was rapidly improving, and in 2020 OpenAI released GPT-3: likely the world's most sophisticated language model to date. The image generation tool Dall-E launched in 2021, followed by Midjourney in 2022. We now live in an unprecedented AI boom: the age of generative AI.

The world was and continues to be fascinated with the possibilities of AI technology. We marvel as AI competes in arenas like chess and trivia. We demand better and better personalized experiences of information. We continue to throw tasks at the AI to think like us, speak like us, drive like us, create like us. We feel a collective sense of magic and awe as we continue to break barriers and benchmarks. With that awe and excitement comes trepidation – along with numerous egregious cases of algorithms gone wrong. But it seems that nothing can slow the exponential boom of technological advancement, and the opportunities that come along with it.

We now sit at a crossroads, perhaps leaving the "move fast and break things" era behind. As Ruha Benjamin suggests, we can opt for "move slow, and empower people" instead [50]. As we adjust to a world centered around information at our fingertips, we now require a vast workforce of adequately prepared professionals to make sense of all the data. Further, we are slowly shifting our sights towards socially responsible and ethically-minded AI. The following chapter lays out the various roles and opportunities in Data Science, and provides overview of how we are currently teaching Data Science in universities across the United States [216, 17, 229, 55]. I introduce leading educational theories in both learning sciences and computing education research, which sets the foundation for my approach to embedding ethics in AI education. I conclude this chapter with my original research on situated learning [286], where I developed a tool for using students' own personal data to teach foundational AI concepts [397]. This work illustrates and ground future discussions of teaching both technical and ethical concepts in parallel and in context of students' lives – illuminating the potential for a more reflective and critical engagement with AI as a whole.

1.2 Demand for Data Science

Before delving in to current trends and context about the increased demand for Data Science and the various roles it entails, it is imperative to briefly clarify some important terms. These

definitions will be used throughout the dissertation and set the stage for the scope of work discussed herein. While there are differing – and sometimes debated – definitions for Data Science and related terms, I will rely in this dissertation on the following conceptualization to differentiate between *Data Science*, *Artificial Intelligence (AI)*, and *Machine Learning (ML)*. Figure 1.1 presents a nested view of these three concepts, and a relational perspective in which Data Science encompasses AI which then encompasses ML. There are other representations of these concepts and their boundaries but I find the nested characterization productive for my work as it allows the term Data Science to serve as the highest umbrella term, encompassing the others as mere components of Data Science. Data Science itself involves the critical thinking, judgment calls, arguments, creativity, and justification needed to solve problems beyond just applying AI algorithms. I will offer some definitions and then summarize the main differences between these fields.

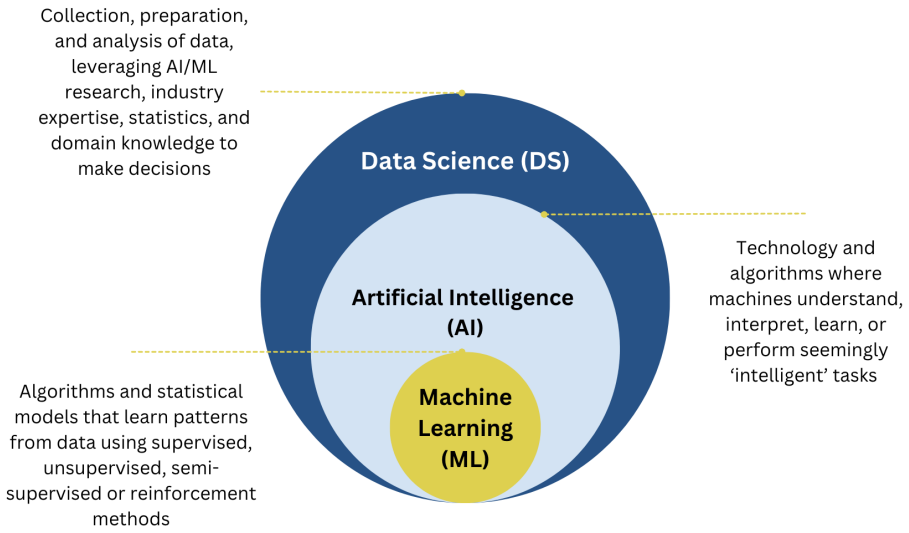


Figure 1.1: A nested representation of the concepts of Data Science, Artificial Intelligence, and Machine Learning. This visualization is inspired from [9].

Definition 1.1

Data Science: Data science combines math and statistics, specialized programming, advanced analytics, artificial intelligence (AI), and machine learning (ML) with specific subject matter expertise to uncover actionable insights hidden in data. These insights can be used to guide decision making and strategic planning[502] in organizations. Data Science encompasses AI and ML but also includes data visualization, data wrangling, data communication, data ethics, and overall analysis of the insights provided by data and models. The role of a Data Scientist includes making decisions about which data and models to use, improving and adjusting those data and models, as well as making decisions based on model output and justifying those decisions.

Definition 1.2

Artificial Intelligence: At its simplest form, artificial intelligence is a field, which combines computer science and robust datasets, to enable problem-solving. It also encompasses sub-fields of machine learning and deep learning, which are frequently mentioned in conjunction with artificial intelligence. These disciplines are comprised of AI algorithms which seek to create expert systems which make predictions or classifications based on input data [501]. Some examples of AI that are *not* considered ML are rule-based systems or solvers that do not rely on data but logic. Classic examples include the N-Queens problem or the Map Coloring Problem, which can be extended to things like scheduling solvers or other constraint satisfaction problems.

Definition 1.3

Machine Learning: Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy, according to IBM [503]. I argue, however, that ML is *not* intended to imitate human learning, though may have drawn inspiration from this metaphor early on. I define ML as a subset of AI that uses data to make predictions, inferences, and/or explain underlying phenomena in the data-generating process. It relies on various models and input to learn patterns in data that can then be used for human decision-making.

While these three concepts are related, the primary differences are that Data Science also includes human decision-making, justification, and *interpretation* from the various models being used. AI encompasses machine learning, but also includes rule-based systems and solvers that do not learn from data. I will rarely, if at all, refer to this kind of AI. In my work, I focus primarily on machine learning algorithms, which are increasingly referred to under the blanket term 'AI'. However, the core skills I am interested in fall under the wider umbrella of Data Science – critically viewing the origins and weaknesses of one's data, making decisions about models and their parameters, drawing insights from model output, and justifying decisions based on those insights and conclusions. In a world where many models are becoming 'plug and play' (or even drag n' drop!), my work is focused on bolstering a Data Science education that prepares future Data Science leaders to think critically, and communicate effectively and compassionately.

In Section 1.1, I traced some the history of the rise of social media alongside our fascination with AI solving human-like intelligent tasks. Today, Data Science and AI have been integrated into almost every sector: social, economic, medical, educational, judicial, creative, manufacturing, agriculture, commercial, etc. The increased demand for Data Science expertise and skills is multi-tiered – the sheer value of data analytics for various industries as well as the global pressures to be the most 'advanced country'.

Across industries, Data Science approaches can be used to make sense of vast amounts of information, generate business insights, automate tasks, recommend products and content, predict outcomes, create media, and even behave and learn like humans (e.g. self-driving cars, intelligent assistants). Furthermore, we exist in a time of a competitive global race between nations to become the world leaders in Artificial Intelligence [320, 446]. The past 30 years has seen an unprecedented leap in AI-driven information technologies, such as search engines, social networking, e-commerce, and streaming services. In the past year alone, we have witnessed a major push to integrate Large Language Models (LLMs) and/or image generation technology into a wide array of services. AI is now an integral part of our digital and daily lives.

The global demand for data-driven services and products is not entirely new. "Data Science" was first coined by Peter Naur in 1974 as "the usefulness of data and data processes derives from

their application in building and handling models of reality" [344]. In 2001, William Cleveland [104] of Bell Laboratories published *Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics*, further reviewed in Section 1.3. In 2006, British mathematician Clive Humby coined the phrase *"data is the new oil"* [27], signalling its great value. This phrase has been built upon, making it clear that data alone is like unrefined and crude oil, and must be processed to be useful [371]. As datasets grew more complex, and more and more of it became available, we entered the era of Big Data: characterized by the 'three V's' of *volume, variety,* and *velocity* [410].

Interest in Big Data – and translating its value into industry returns – continued to grow. In 2011, the McKinsey Global Institute published a report entitled *Big data: The next frontier for innovation, competition, and productivity* [304]. This was an influential report that set the stage for the development of Data Science programs in the following years. The report states:

> *"Leaders in every sector will have to grapple with the implications of big data, not just a few data-oriented managers. The increasing volume and detail of information captured by enterprises, the rise of multimedia, social media, and the Internet of Things will fuel exponential growth in data for the foreseeable future."*

Also setting the stage for the various Data Science educational programs to follow, the report claims:

> *"There will be a shortage of talent necessary for organizations to take advantage of big data. By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions."*

With the immense value of making sense of data across nearly every domain, combined with this projected shortage,*Data Science* became the next big thing – an inescapable buzzword. This section would be remiss without mentioning the 2012 Harvard Business Review article that referred to Data Science as *"the sexiest job of the 21st century"* [127]. Nearly every media article about the growth of Data Science references this viral piece, academics also took notice [77]. The article itself described the increased demand for "a high-ranking professional with the training and curiosity to make discoveries in the world of big data". They describe some necessary skills: coding, curiosity, the ability to conduct experiments and ask the right questions. They mention increased demand in industry, and the start of some undergraduate Data Science programs across the US. They point to Big Data as an "epic wave starting to crest". The article content is an important documentation of the current trends, describing the widespread frenzy of demand for specialists with Data Science skills. However, I argue that the terminology and framing of "sexiest job" is a detriment – not only can sexualized terminology be inappropriate coming from a male-dominated field, but it also may give the connotation that the job is only about prestige and money. It may even contribute to overconfidence and ignorance to complex and ethical dilemmas in the field. The moniker was meant to represent the rising demands and opportunities for Data Science skills, which likely could have been accomplished with *"the most exciting job of the 21st century"*.

Headlines aside, the demand for Data Science has steadily risen since the 2012 article [470]. The widespread adoption of social media may have been a large factor in such demand. For example, in 2006 LinkedIn discovered that by harnessing network analytics, they could increase clickthrough rate by 30% by recommending 'People You May Know', increasing engagement on their site [127]. Social media came along with massive amounts of personal data, which could be used to predict, and manipulate, human behavior. Banking, e-commerce, healthcare,

and global research initiatives were all waking up to the promise of Big Data analytics [504, 304]. This led to increased job opportunities, with a 2019 Indeed article showing that Data Science job listings rose 256% since 2013 [126]. A 10-year follow up by the Harvard Business Review confirmed that Data Science continues to be in demand, with the addition of increased interest in Data Science ethics, algorithmic bias, responsible AI, and transparency. The new report also refers to the 2016 Cambridge Analytica scandal, where personal data was used to influence voters in the 2016 election, as a turning point for the increased demand for regulation and oversight of potentially harmful data practices [244].

This specific interest in responsible AI was hinted at early on, with the 2011 McKinsey Global Institute report [304] stating:

> *"Several issues will have to be addressed to capture the full potential of big data. Policies related to privacy, security, intellectual property, and even liability will need to be addressed in a big data world. Organizations need not only to put the right talent and technology in place but also structure workflows and incentives to optimize the use of big data. Access to data is critical – companies will increasingly need to integrate information from multiple data sources, often from third parties, and the incentives have to be in place to enable this."*

Today, we sit in a time of opportunity: AI continues to grow, and so do calls for its ethical and responsible use [282, 6]. In order to meet the rising demand, we must sufficiently train future Data Scientists – they must be competent coders, statistical thinkers, experimental scientists, strong communicators, and ethically minded practitioners. How do we do it? The next section covers various efforts in Data Science education, pointing to the widespread required skills to succeed as a Data Scientist and demonstrating the need for a better foundation in ethical thinking in Data Science curricula.

1.3 Data Science Curricula

In the past few decades, different approaches to teaching Data Science have been proposed, with debate over which topics should take precedence and what components delineate Data Science from other subjects like Statistics or Computer Science. While the term 'Data Science' is as heavily debated, I provide a definition of the Data Science practitioner, the most in-demand job of the 21st century, but a profession that dates back to the very beginnings of statistics, mathematics, and computing:

Definition 1.4

Data Scientist: an analytics professional who is responsible for collecting, analyzing, and interpreting data to help drive decision-making in an organization. The role of Data Scientist includes making decisions about which models and data to use, improving and adjusting those models and data, as well as making decisions from the output and justifying those decisions. They may not be directly responsible for all processes in the data science lifecycle but have a breadth of knowledge that allow them to guide decision-making and recommendations for what data and algorithms are useful or required (IBM).

Over the past decade, "Data Scientist" has since evolved into a clearer, though still debated, role, comprising of a variety of necessary and core skills, as well as optional, specialized competencies. Over the past decade, more Data Science degree programs have been launched across the U.S. These consist primarily of Master's level programs and then Doctoral degrees [453], with few undergraduate programs [17]. This has occurred in tandem with increased Computer

Science education across K-12, with AI as a component [475]. Computer Science, Statistics, Engineering, Psychology, Life Sciences, Business, and Information Science departments also provide an increasing number of of Data Science courses, though AI and Machine Learning have historically strictly 'belonged' to Computer Science and Statistics departments.

An early action plan for Data Science curricula came from Cleveland [104] in 2001, at the time referring to Data Science as a way to "enlarge the technical areas of statistics" for the data analyst. The plan lays out six technical areas of work for a university department along with their respective percentage of resources to be allotted. The six areas were: 1) Multidisciplinary Investigations (25%), 2) Models and Methods for Data (20%, 3) Computing with Data (15%), 4) Pedagogy (15%), 5) Tool Evaluation (5%), and 6) Theory (20%). Of note is the early foresight that a core component of a Data Science curricula ought to focus on pedagogy itself, stating:

> *"A data science department in a university must, of course, concern itself with teaching in its own setting. But it is vital that resources be spent to study pedagogy and to teach pedagogy. It makes sense that such study encompass more than the university setting; curricula in elementary and secondary schools, company training programs, and continuing education programs are important as well. Education in data science does many things. It trains statisticians. But just as important it trains non-statisticians, conveying how valuable data science is for learning about the world." [104]*

By the 2010's, and following the Harvard Business Review article in 2012 [244], higher education "Data Science" programs had been launched across the nation. Amongst reviews of Data Science curricula, we see focus on topics of data workflow, algorithms, optimization, statistical testing, data visualization, interpretation, and communication. Typical for the early development of Data Science curricula, 'ethics' is either left to the end or left out entirely [216, 55, 147]. This may be due to the fact that high profile cases of AI bias and algorithmic harm were not yet as widely acknowledged as they are today (*e.g. Buolamwini and Gebru [80]'s* Gender Shades *comes out in 2018).*

In a 2014 design for an undergraduate degree in Data Science, Anderson et al. [17] proposes that the curriculum ought to include 8 key topics:

1. **Large data sets / streams:** create/design, access, clean, analyze, aggregate, organize, visualize
2. **Database:** design, storage, query, modeling
3. **AI techniques:** genetic algorithms, neural networks, Bayesian networks, intelligent agents, machine learning, pattern matching, heuristic search, knowledge representation and ontologies, natural language processing
4. **Software and Algorithms:** design, programming, testing, algorithms and analysis
5. **Information retrieval:** information theory, data mining, text mining, image mining, indexing, content analysis, linguistic processing, abstracting, search and retrieval, information filtering, query formulation
6. **Mathematics:** logic and counting, discrete structures, statistics, linear algebra, modeling and simulation
7. **Oral and written communication:** effective communication

8. **Social, ethical and legal issues:** privacy and security, property, policy, information validation, professionalism

– An Undergraduate Degree in Data Science: Curriculum and a Decade of Implementation Experience [17]

Other curricula reviewed contain similar elements, differentiating Data Science from traditional statistics by focusing on the computational and interpretive elements of what it means to 'think with data' [216]. Several of the calls for improved Data Science education include a focus on real-world cases, with De Veaux et al. [129] stating:

> *"The recursive data cycle of obtaining, wrangling, curating, managing and processing data, exploring data, defining questions, performing analyses, and communicating the results lies at the core of the data science experience ... Students of data science must encounter frequent project-based, real-world applications with real data to complement the foundational algorithms and models."*
>
> *– Curriculum Guidelines for Undergraduate Programs in Data Science*

Another similar on real-world cases comes in the form of what Hicks and Irizarry [229] refers to as *computing, connecting, and creating*. The authors describe how traditional Statistics courses are missing a focus on **computing**, which Data Science centers around. Further, they describe how traditional Statistics courses present abstract mathematical notation as opposed to real-world cases and data problems, and that **connecting** with the data and cases is key for success. This notion of **connecting** can be directly mapped to the concept of situated learning [286], a learning theory that I rely on in my work on AI education [397]. Situated learning will discussed in more detail in Section1.4. Hicks and Irizarry [229] also encourage a focus on students **creating** their own research questions and interventions with data as a core part of the curricula. They reinforce the iterative data science pipeline of *Problem, Plan, Data, Analysis, and Conclusion*, which can be used to develop Data Science curricula today. Importantly, their curriculum design neglects to include any social impacts or ethical responsibilities in Data Science. It is only in recent years that we have seen a notable push towards Data Science ethics, AI ethics, Responsible AI, and Algorithmic Fairness [116, 6]. As Fiesler, Garrett, and Beard [169] point out, a growing number of CS instructors are including ethics as part of their courses, and are not "asleep at the wheel" as one 2017 article claimed when it came to ethics in CS education. They write:

> "This reminder – that code is power, and it should be used responsibility–could be part of every computing course, but is arguably most important at the very beginning of the process of learning to code. This strategy might even be a way to combat an 'I'm just an engineer' mindset that ethics is 'someone else's job' by emphasizing its role in computing from day one and then continuing this reminder throughout the curriculum."

I elaborate on these efforts in Chapter 3, but effective strategies for teaching AI ethics are still nascent, with this dissertation attempting to contribute to this conversation by offering strategies for embedding ethics into Data Science and AI curricula. In order to understand how to teach *ethics* in AI, it is useful to ground the approach in fundamental theories of how to teach in the first place. The following section introduces some of these foundational theories, providing an overview of learning theories such as *constructionism*, *experiential learning*, *situated learning*, and how we can apply them to AI education specifically.

1.4 Theoretical Foundations & Learning Sciences

One proposed approach to teach both technical and ethical concepts is to ground a student's learning in their own experiences and identities. I follow this idea throughout this dissertation, providing different ways of infusing one's personal experiences with unjust algorithms as a learning opportunity. However, before digging into this opportunity in detail, I first introduce some of the theoretical foundations of learning sciences and education theory. What do we know about best practices in teaching? We ask questions such as: *how does one impart knowledge to another?* You've likely heard the phrase "give a man a fish and he'll eat for a day, teach a man to fish and he'll eat for a lifetime". But how does the fisherman know how to teach? There is a long history of work in psychology and education that investigate and describes what we know about learning and teaching. According to the early Behaviorist theories of the late 1800s, punishments and rewards could be used to teach – the fisherman might use electric shocks when the pupil does something wrong, and rewards when the pupil does something right. After a long enough time, the pupil could be trained to do the proper series of steps to catch a fish [439, 206]. This classical conditioning, according to Pavlov, may result in the pupil being excited to eat fish simply by being shown a fishing pole after enough repetitive training. Luckily, ideas around education have progressed much further a conditioned response based on a stimulus.

Following Behaviorist educational theory came Cognitivism: the study of introspection, mental structures, and how learners acquire and retrieve knowledge when being taught [224]. In our case, we would be curious about what new information our pupil was learning about fish, bait, equipment, and other tips the fisherman would provide. The pupil would be developing a kind of schema, such as "when I go to this location with this kind of bait and wait this amount of time, a certain type of fish may come along and bite my hook, which I then reel in to catch the fish". Cognitivism is interested in the mental processes going on in the pupil's mind; separate from their behavior. "Executive functioning" is the set of skills which allow for planning, organization, metacognition, self-control, working memory, and adaptable thinking. Cognitivism is particularly interested in these skills, and later developed into Cognitive Psychology which based mental processes in theories of computation [425].

Beyond executive function, human cognition also involves social reasoning; we look to others to know how to act, especially in childhood. Children will imitate behaviors of others that are rewarded, and abandon behaviors that seem to be punished. Part of their mental model is that some ways of doing things are more beneficial than others, and they can rely on social data to figure out which. Bandura introduced social learning theory which includes physical demonstrations of behavior for the learner to watch [33]. In this case, the pupil watches how the fisherman does each step in the fishing process and tries to imitate what works well.

So far, all of these theories seem like feeding data into a predictable box. Do something enough times, create the right mental structures, see how it's done, and you're bound to succeed. But none of these theories take into account how individual experience shapes what we learn. Each individual learner comes in with their own expertise, culture, language, viewpoints, connections, and identity. This individual experience is not limited to the cognitive, either. The way that the learner's physical body moves through space affects how they experience the world, and which strategies will work best for them for a given task. This doesn't mean that education should be perfectly crafted for each individual experience so much so that everyone is learning something completely different. It means that expression of self, experience, and culture should be utilized in the education process. This notion is core to the next movement in educational theory: Constructivism.

Swiss psychologist Piaget asserted that the learner comes in with experiential knowledge of their own, even if they are a very young child [255]. Even three-year-olds have begun to develop identity, preferences, personality, fears, and associations (ask any parent!). Constructivism

challenges the idea that there is a "blank slate" learner, waiting to be trained. Instead, there is a process of co-constructing knowledge alongside experience and individuality. Vygotsky's "Zone of Proximal Development" would further argue that there is a gap between what a learner knows and how much facilitation they need from others in order to gain certain kinds of knowledge [76]. In other words, they still need a teacher, even if they are constructing their own knowledge through experience. Similar to the academic trying to answer research questions, they *might* come across the answers based on their own experience, but progress would be painfully slow without reading the successes and failures of others.

Within the field of computing education, Seymour Papert is most famous for applying constructivist ideas to educational methods via construct*ion*ism [374, 217, 11]. The idea here is to allow learners to construct their own physical or digital artifacts that they, or others, might actually use. While construct*iv*ism focuses on how learners create their own mental structures via experiencing, construct*ion*ism gives them a vessel to do so. Seymour Papert's learning theory of *constructionism* [374, 376, 217] significantly shaped the way we teach and learn computing.

Definition 1.5

Constructionism: A learning theory developed by Seymour Papert [374] which advocates for student-centered, discovery learning where students use what they already know, to acquire more knowledge. It posits that learning results from exploration, discovery, collaboration, participation in authentic activities, and has varied and unique outcomes for each learner[10]. It builds upon Piaget and Vygotsky's *Constructivism* [16], with a focus on a learner's construction of some kind of public artifact: "whether a sand castle on the beach or a theory of the universe." [377]

Constructionism is a well known theory in the field of computing education, lauded by MIT Media Lab and Scratch founder Mitchel Resnick [402] as "fundamentally changing the way we think about learning, the way we think about children, and the way we think about technology". Scratch is a block-based programming language primarily for children, who use the drag and drop functionality of different programming pieces to build their code and their creations. This draws from the LOGO programming language developed by Seymour Papert, Wallace Feurzeig, Daniel Bobrow, and Cynthia Solomon, a learning environment where children explore mathematical ideas and create projects of their own design, using a graphical or robotic turtle [449]. This was the first programming language directly aimed at children's learning, and served as a foundation for computing education today [114].

The ideas of constructionism make sense for a child learning why ice melts or how red and blue make purple. But it gets less clear when we want to teach about complex societal issues. How do we build artifacts that help students engage in issues of race, gender, poverty, health, or crime? Therefore, I also turn to situated learning and situated knowledges – related but different learning theories that explore how people rely on their identities, perspectives, and lived experiences to learn.

Definition 1.6

Situated Learning: Learning that takes place through the relationships between people; connecting prior knowledge with authentic, informal, and often unintended contextual learning; an instructional approach developed by Jean Lave and Etienne Wenger in the early 1990s [286], and follows the work of Dewey [139], Vygotsky [76], and Clancey [103], who argue that students are more inclined to learn by actively participating in the learning experience. Situated learning essentially is a matter of creating meaning from the real activities of daily living where learning occurs relative to the teaching environment. "Learning, thinking, and knowing are relations among people engaged in activity in, with, and arising from, the socially and culturally structured world." [286]

Situated learning specifically focuses on the idea that "learners inevitably participate in communities of practitioners and that the mastery of knowledge and skill requires newcomers to move toward full participation in the sociocultural practice of a community. A person's intentions to learn are engaged and the meaning of learning is configured through the process of becoming a full participant in a sociocultural practice" [286]. By learning *in context*, learners see the applications of their knowledge and move into the roles of community members and eventual experts. My paper "*Learning machine learning with personal data helps stakeholders ground advocacy arguments in model mechanics*" [397] was motivated by both Constructionism and Situated Learning to explore more effective ways of teaching AI/Machine Learning concepts, with positive results. Later work has further bolstered these claims that situated learning is beneficial for teaching AI [431, 256, 373] with various successful efforts to embed student learning in their own experiences and contexts. In particular, we can approach both technical concepts and *ethical reasoning* using a situated learning approach.

Haraway [214]'s *Situated Knowledges* goes further to bring in the dimensions of oppression and marginalized identity and how learners possess unique knowledge from a marginalized standpoint. Because the topics of AI ethics deal with such complex and social issues, we must bring in feminist theory of learning to properly support learners engaging with difficult cases of algorithmic harm, which I introduce in a variety of the published works in this dissertation – theories such as Situated Knowledges, feminist phenomenology, counterstories, trauma-informed computing, and ethics of care inform much of my work.

In a reduced form, Situated Knowledges can be described as rejecting 'objectivity' in knowledge production, and acknowledging how power relations shape one's abilities and perspectives. There will always be subjective interpretation of truth, shaped through one's identities, privileges, oppression, and lived experience. It is the case that marginalized populations bring unique and valuable information to knowledge production, and that their lived experiences allow them to see what may otherwise be overlooked. As mentioned in the previous sections, the past decade has seen more and more marginalized populations speaking on the ethical concerns regarding AI technology. It was due to Joy Buolomwini's identities as a Black woman and poet that allowed her to so deeply explore the impacts of racially biased facial recognition technology in the 2018 paper *Gender Shades*[80]. We see the importance and *value* of diverse perspectives from scholars like Safiya Noble [352], Ruha Benjamin [50], Timnit Gebru [47, 80], Deborah Raji [394, 396], Emily Bender [47], Catherine D'Dignazio, Lauren Klein [116], Margaret Mitchell [326], Cathy O'Neil[360] and so many others.

The call for socially responsible and ethically minded AI is here, with opportunities to improve Responsible AI and AI Fairness appearing every day. It is my belief, founded in both my research and the research of others, that to properly prepare future Data Scientists for ethically-minded careers, we must allocate resources to computing education research and curriculum development with embedded ethics alongside technical instruction. In our 2020 piece *It Is Time for More Critical CS Education* we emphasize that studying algorithmic harms

is not enough.

> *" It means more than just an ethics requirement for CS majors: it means recasting computing itself in moral, ethical, and social terms. Realizing a more critical CS education requires more than just teachers: it also requires CS education research. How can we convince students they are responsible for what they create? How can we make visible the immense power and potential for data harm, when at first glance it appears to be so inert?" [275]*

Directly using what we discovered in the research contained within this dissertation, Ko et al. [274]'s *Critically Conscious Computing* textbook outlines strategies for teaching technical AI concepts alongside ethical critique of AI algorithms, in a situated way using learners' own data. The following work serves as the jumping off point for embedding ethics in technical AI/ML education, relying on students' own perspectives, experiences, and data. Register and Ko [397] describes the benefits of relying on personal data to learn about AI and Machine Learning. I describe how to utilize the students' own *funds of knowledge* [1] in an advocacy task resisting algorithmic harms. We discover that students who used their own personal data had a boost in both technical and advocacy skills in an AI tutorial.

[1] "historically accumulated and culturally developed bodies of knowledge and skills essential for household or individual functioning and well-being"[329].

1.5 Situated Learning on Personal Data

1.5.1 Author's Preface

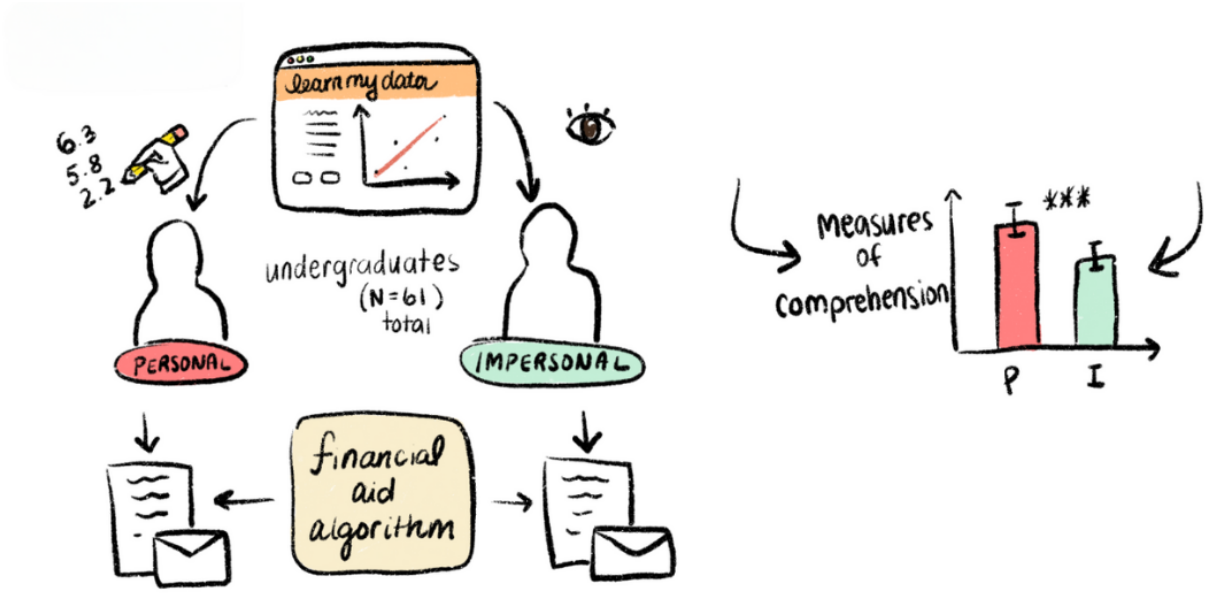


Figure 1.2: Original artwork depicting the study design and cartoon representation of results. Participants in different groups went through a machine learning tutorial, followed by an advocacy task, where their advocacy letters were analyzed for comprehension.

Throughout my Data Science education, I had noticed a gap between how it was being *taught* and how I *used it* in my daily life. I was surrounded by algorithms and data, and yet my education only reflected a small portion of that – relying more on outdated datasets from the 1970's or only presenting complicated math proofs. As I trudged through machine learning concepts, I started to see them everywhere. I was also starting this work during an increasingly loud and pervasive hum about AI, oppression, racial justice, and algorithmic bias. *Weapons of Math Destruction*[360] was published in 2016. *Gender Shades*[80] was published in 2018. *Algorithms of Oppression* [352] also hit the scene in 2018, quickly followed by *Race After Technology* [50] and *Data Feminism* [116]. Although many people had been working in AI ethics long before I was, it also felt like at the start of my PhD I was riding a wave of these important and timely issues. Alongside my own exploration of identity and advocacy, I knew I wanted my work to combine education with activism.

I wanted to jump in right away to get people to look at their own social media data – the most pervasive machine learning-based system I knew. I also wanted to start simple – demonstrating how linear regression is often a relatable and foundational step into the world of more complicated ML. I piloted my design with undergraduates, having them plot out the relationship between the number of comments and the number of likes they got on their Instagram posts. This proved interesting, as all of a sudden non-experts had sophisticated things to say about why or why not a post would succeed in gaining attention and engagement. I witnessed everyday users of ML technology turn into analyzers of their own data. I was captivated by the promise of using personal data to understand both machine learning processes and the impacts they have on us.

I ultimately chose to pivot away from social media data and instead explore one of the examples from *Weapons of Math Destruction*. At the time, I didn't have enough understanding of how machine learning related to social media, or even the potentials for harm in those systems. I relied on a more academically relevant case study instead: a vastly simplified version of admission metrics. I had students rate their interest in their courses and the grades they got,

and learn about linear regression through the correlation between the two. They had to reduce a complex phenomena like "interest" into a 1-7 scale, as well as respond to a scenario in which this poor metric was used as a proxy for their grade. While it was certainly a toy scenario, it allowed for a good teaching tool. Next, I introduced them to a financial aid scenario, where the stakes were higher. In the scenario, the admissions office used number of children in a family as proxy for predicting Financial Aid amount awarded. I liked this example, as it was a more highstakes and realistic case of prediction; it had a lot to comment on by participants in the study. It was personal and relevant to students, while also having enough issues that they could write empowered arguments. In hindsight, I would have directly used the admissions metrics example as opposed to Interest vs. Grades (an unlikely scenario). Using an admission metrics example alongside the financial aid example would allow students to think critically about how they might be reduced down to numbers and analyzed for future success. It also may have been depressing. Overall, I was simply getting my foot in the door thinking about situated learning and a more personalized way of teaching machine learning concepts. I was especially excited about using self-advocacy as a metric of comprehension and success. Drawing on ideas of patient self-advocacy in hospital settings [219] as well as disability justice [268, 440, 195, 194], I wanted to formulate ideas around what self-advocacy looked like in machine learning contexts. As you will see throughout this dissertation, self-advocacy alone is not enough. Challenging and repairing AI bias and algorithmic harm will be a vast, multi-faceted, and coordinated endeavor far beyond the scope of what one researcher can accomplish. But this first publication of mine was the small seed that set the stage for the rest of my work – with situated learning and self-advocacy at the forefront.

Verbatim Text

This marker denotes the beginning of previously published work, reference below.

Register, Yim, and Amy J. Ko. (2020). "Learning machine learning with personal data helps stakeholders ground advocacy arguments in model mechanics." *Proceedings of the 2020 ACM Conference on International Computing Education Research.*

Learning Machine Learning with Personal Data Helps Stakeholders Ground Advocacy Arguments in Model Mechanics

Abstract

Machine learning systems are increasingly a part of everyday life, and often used to make critical and possibly harmful decisions that affect stakeholders of the models. Those affected need enough literacy to advocate for themselves when models make mistakes. To understand how to develop this literacy, this paper investigates three ways to teach ML concepts, using linear regression and gradient descent as an introduction to ML foundations. Those three ways include a basic *Facts condition*, mirroring a presentation or brochure about ML, an *Impersonal condition* which teaches ML using some hypothetical individual's data, and a *Personal condition* which teaches ML on the learner's own data in context. Next, we evaluated the effects on learners' ability to self-advocate against harmful ML models. Learners wrote hypothetical letters against poorly performing ML systems that may affect them in real-world scenarios. This study discovered that having learners learn about ML foundations with their own personal data resulted in learners better grounding their self-advocacy arguments in the mechanisms of machine learning when critiquing models in the world.

Introduction

Machine Learning systems are increasingly becoming a part of everyday contexts, such as medicine, finance, legal decisions, transportation, social media, entertainment and more. While many people think of "Artificial Intelligence" as *The Terminator* [97, 487, 66, 82], or something that will "take over the world" [528], they may not be recognizing that Machine Learning (ML) technology is integrated into almost everything we do on our phones, including Google Maps, Facebook, text prediction, face recognition, photo tagging, friend and media recommendations, spam detection, and information retrieval in search engines.

There are potentially harmful effects of ML systems, beyond just a poor facial recognition in a Snapchat application trying to detect your face. Many ML systems are used to determine whether people can get a loan, buy a house, receive government assistance, be eligible for employment, are likely to have committed a crime, will successfully reintegrate into society after incarceration, and many other impactful decisions about human lives [121, 270, 75, 513, 279].

Stakeholders of ML systems should feel empowered to speak up against models that affect them, yet few people can actually explain how these systems work. Contrary to preconceptions, ML is not just for computing majors, as ML spans topics from astrophysics to zoology [496, 497]. For non-experts to advocate for themselves in ML scenarios, they should be able to reason about whether or not a tumor-detection system is trustworthy, knowing which political news they will see on their newsfeeds and why, or how to interpret the recommendations from a

prisoner recidivism predictive model. Widespread ML literacy would be important for jurors, consumers, voters, policymakers, engineers, designers, journalists, and more.

Core features of this literacy include model transparency, understanding the mechanisms, contextualizing data, critical thinking, and leveraging learners' interests and backgrounds [295]. This would allow for agency and more targeted self-advocacy for those affected by ML. This means that someone would be able to articulate the flaws in the design of various ML systems, be able to ask effective questions, and be able to express critiques and solutions for their own interactions with ML. This kind of self-advocacy within the machine learning domain might look like:

- Jurors asking questions about a predictive tool used to argue if an incarcerated individual should be granted parole, and discussing if it was fatally biased to favor past judicial decisions.
- Patients asking a doctor if a computer vision tool was trained on their particular condition, and asking about common mistakes the model has and how it is accounted for.
- Voters having basic knowledge about why an article appeared on their social media, and how newsfeeds can be biased by what they and their friends click on or "Like".
- Loan borrowers asking creditors what features were included in models determining whether they should be approved for a loan, and what data that model was trained on.
- Potential employees questioning a company using an NLP model to filter resumes for hire, pointing out that their name or writing style may influence the hiring decision due to biased training data.

Being able to make these critical judgments about ML systems likely relies on the ability for stakeholders to understand the mechanisms of those systems; the inputs and outputs, the strength of the relationship, the appropriateness of the features used in prediction, the shape of the model being used, and the fit of the model to the data. Each of these skills would provide stakeholders with necessary insight to resist against models making incorrect predictions and allow them to identify alternatives or fatal flaws in those ML systems. This literacy is not at the level of programming or innovating on the systems themselves, but it is more generalizable than simply knowing facts about ML systems [222, 115, 387, 131, 424].

Resources for teaching ML often teach specific programming techniques and code libraries for specific problems that are often irrelevant to learners personally, such as: "predicting price of diamonds by hardness, predicting type of iris by length of its petals, predicting survival rates on the Titanic". The most well-known Machine Learning course is Andrew Ng's Coursera course [349], which is mathematically heavy and also relies on the common datasets in the ML community. Similarly, data science education research includes experience reports of experimental data science courses, or has contributed tools to help learners explore their data [32, 280, 235, 234, 18, 39, 38], but without scaffolding and real-world context combined. These also focus on university-level instruction or those with background already, and rely on having weeks to months of material to situate the learner. AI literacy for children often involves physical machines such as robots or voice assistants, and doesn't explicitly focus on tools for resistance against harmful models [532, 234, 149, 264, 474, 125]. Developing ML literacy demands different techniques than the traditional lecture model if it is to reach diverse populations [54, 119, 120, 380]. None of the existing approaches directly teach the ML literacy necessary for stakeholders to make critical judgements about the ML systems impacting their daily lives, and we do not yet know if the current resources generalize to helping learners understand ML systems in real-world personal scenarios.

One way to teach machine learning literacy for self-advocacy is to link the mechanisms of ML to the learner's own prior knowledge and experience [380, 131, 32]. To do this, we could incorporate learners' funds of knowledge: leveraging the learner's already existing knowledge and experience by strategically teaching material revolving around the learner's culture, situated knowledge, and relationships [329, 192, 193], a concept adjacent to Papert's constructionism [374, 375] and Dewey's experiential learning [139]. There are some projects that have incorporated relevant and interesting data into data science education, such as CORGIS (Collection of Really Great and Interesting dataSets) [38]. However, there is greater potential for integrating personal data and experience into teaching data science, especially for situated and justice-oriented projects [507, 309]. Because ML systems are so intricately tied to the data they process, we theorize that integrating personal experience and domain knowledge into the teaching of ML literacy could benefit learner's understanding of the mechanisms at play, and teaching with the learner's *own* data allows them to situate themselves with regards to the ML system. Using personal data could increase learners' attention on the mechanisms of ML by making the mechanisms more personally interesting; this increased attention would lead to 1) better understanding of the mechanisms, 2) better ability to apply the mechanisms of ML to their lives, 3) critiques of ML scenarios that are more explicitly grounded in the mechanisms of ML.

To test if leveraging learners' personal funds of knowledge benefits their ability to self-advocate against potentially harmful ML systems, we designed three forms of instruction and an empirical study to evaluate them. In order to give learners the best chance at learning about the mechanisms of ML, we designed a tutorial using best practices from learning sciences to teach a foundational ML concept: linear regression with gradient descent. We used that tutorial to teach using the learner's own personal data (*Personal condition*) vs. a hypothetical individual's data (*Impersonal condition*). We compare these conditions to a baseline description of facts about ML systems and linear regression without referring to any data (*Facts condition*). We studied the impact of those interventions on learners' ability to self-advocate in real world ML scenarios by asking learners to write a letter to the enforcer of a model that made a wrong prediction. This mimics a reasonable pathway in the world; speaking up for yourself in medical, financial, digital, legal, and institutional scenarios.

Possible Ways of Teaching Machine Learning Literacy

It is unclear how to effectively teach stakeholders to make critical judgments of ML systems, but we can go through some possibilities drawn from what we see in current ML education. First, we might try to teach prediction by providing closed-form mathematical equations, which are often used to teach ML at the university level. These are likely not comprehensible by the average person using ML systems because they require a lot of math background. Moreover, even if the learner did understand the equation, they would still rely on further mental simulation to determine effects of the model on different kinds of data. Equation 1 demonstrates how knowledge of sigmas, subscripts, weights, vectors, and more would be required to reason about linear regression from the closed form representation.

$$y = w_0 x_0 + w_1 + x_1 + ... + w_m x_m = \sum_{j=0}^{m} w^T x \tag{1}$$

Instead of mathematical explanations, we could teach ML literacy by providing the general facts about ML systems and how they work, similar to a presentation about ML, but this could be insufficient for learners to trace how new or unusual data might be manipulated by the system

or how it might play out for them personally. For example, a consultant might tell a client that ML suffers from "garbage in, garbage out." The client may now rightfully distrust models with flawed data, but will still have to further mentally simulate to understand the severity of the consequences, or to offer alternative solutions that work better. Given facts about ML systems, they may be unable to ground those general facts in how it applies to them personally or where such systems are used in the world. They may also not have enough information to reason out alternative ways the model could work or be able to pinpoint what kinds of data cause specific models to fail. All of these skills are useful for successful self-advocacy when critiquing an ML model.

While the above techniques are less amenable to supporting the learner to make critical judgments, we could teach the idea that ML algorithms are responsive processes that manipulate data by describing the steps of the algorithmic process. This might give the learner more insight into where the model can fail. Consider this explanation of Gradient Descent:

> "Gradient Descent works by starting with random values for each coefficient. The sum of the squared errors are calculated for each pair of input and output values. A learning rate is used as a scale factor and the coefficients are updated in the direction towards minimizing the error. The process is repeated until a minimum sum squared error is achieved or no further improvement is possible."

However, describing the algorithmic process without any data does not say anything about how the algorithm would respond to new scenarios and is removed from relevant context. This could result in the learner ignoring crucial data scenarios that would happen in the real world, such as models missing outliers or not accounting for important features for a problem. To use machine learning vocabulary, the learner might "underfit"

Using actual data to trace through a problem could give more insight into how that data gets manipulated and where the system may fail. One promising way to teach about ML systems would be to describe an algorithmic process by actually following a trace of some data, demonstrating how data is manipulated and can affect the outcome (as in prior work that explains programming language semantics [347, 517]). However, typical datasets used are either completely abstract ("Product A, B, C and x, y, z") or irrelevant and unapproachable to the learner; who might lack context or domain expertise about the example dataset. For instance, the common iris dataset includes variables like "sepal length, sepal width, petal width and petal length" to predict the various species of iris (*setosa* or *versicolor*). Without context or domain knowledge, the learner may not have any intuition about what is correct or incorrect in their model, or where the model fails. They may trust the results of a faulty model due to lack of insight into the data. Furthermore, such "toy projects" do not engage with societal impact [225].

One way to ensure that the learner thinks more critically about the ML system they are learning about is to fully immerse them in the data process [526]. By using their own funds of knowledge and their own data, the learner must automatically grapple with the nuances of the algorithmic process from start to finish. We theorize that leveraging personal data is particularly suitable for teaching ML literacy because the outputs of ML systems rely critically on relationships in data, and allowing the learner to draw on their own knowledge of the data domain could contribute to better understanding of how ML mechanisms relate to them in the world [477, 343, 113]. Drawn from Dewey's *Experience and Education*, we theorize that creating an experience that situates the learner in the data domain may allow the learner to more readily construct a basis for understanding how ML is working on that data [139]. Using personal data automatically means using different data for each individual learner. This means that the learner may get the chance to explore "dirty" data, or data that is not compatible with the model they are learning about. This may prompt them to think about the pitfalls of the ML techniques in general, which could strengthen their self-advocacy arguments. Instead of learning about

algorithmic bias in theory, this technique allows learners to confront how algorithmic bias may affect their own data. We theorize that it might provide the learner with agency to explore their own biases towards what is "objective" in data science, while also giving them richer insight into possible solutions against algorithmic bias. We know that higher level design decisions are some of the most difficult ML concepts to teach [462, 461], and this work demonstrates that integrating personal data and self-advocacy tasks may prompt learners to engage with those tasks in a natural way.

While this seems promising, it may also be the case that learners "overfit" to their own experiences, and are unable to think of other people or scenarios affected by the models. It could be the case that they hyperfocus on their own data, without considering the average use cases for the model. Given that we want to test the effect of using personal data on learner's engagement with the ML tutorials, we arrived at the following research questions:

- **RQ1: Do learners using personal data pay more attention to the mechanisms of machine learning?** We theorize that using personal data would be more interesting and relevant to the learner; resulting in them paying more attention to the tutorial they were given. In particular, they would pay attention to the actual mechanisms of machine learning; and refer to more of those mechanisms in their critiques and self-advocacy arguments.
- **RQ2: Do learners using personal data have a better ability to apply the mechanism to their life?** We theorize that using personal data would allow the learner to draw on relevant domain knowledge from their own experiences; we explore if learners reference their personal experiences more if they use their own data to learn about linear regression.
- **RQ3: Do learners using personal data ground their self-advocacy arguments in the mechanisms of machine learning?** The ability to self-advocate against potentially harmful machine learning models relies on being able to articulate critiques of the model at hand. Successful self-advocacy relies on articulation, negotiation, domain knowledge, and problem solving skills. We look for evidence of these skills in relation to the machine learning scenarios. We explore how using personal data to learn ML relates to learners' self-advocacy arguments.

LearnMyData Tutorial

In order to test how learning ML on personal data affects the ability to critique ML systems and self-advocate, we needed to create a custom tutorial that took in personal data as the data used to teach the mechanisms of machine learning. We decided to teach univariate linear regression (one predictor variable and one response variable) and gradient descent as a proxy for other introductory machine learning concepts. The *LearnMyData* tutorial improved upon the most popular linear regression tutorial for introductory Machine Learning: Andrew Ng's Coursera course videos. We did this by combining some content from the original Coursera material and by introducing promising design practices from Learning Sciences, including but not limited to: minimal visual design, engaging the learner by asking for feedback along the way, drawing upon knowledge that the learner already has, and designing for self-paced learning. The learning objectives of the tutorial were to describe univariate linear regression, communicate how machine learning "fits" a model to data in order to predict new data, and how the model must be "trained" on data that may or may not generalize. We used the *LearnMyData* tool for both the personal data instructional design and impersonal data instructional design, with the former including an input table for learners to input their own data, and the latter framed around

some hypothetical data. To present ML facts to the learners, they did not see the *LearnMyData* tool, but instead got a printout sheet of similar content (with one hypothetical example about grades) without interactivity.

Figure 1.3: Selected screens from *LearnMyData* tutorial, demonstrating some of the interactive elements of the tool. Each screen is accompanied by a paragraph of text with instructions (not shown). The **Collect Data** (1) screen is unique to the *Personal condition*, where the learner inputs their own data. The **Draw a Line** (2) screen gives the learner a chance to try to draw their own best fitting line. Later on they see the residual values on the true best fitting line (3), and reason about the parameters of the model in terms of $y = mx + b$ (4). They watch a poor way of determining those parameters, which would be to **Guess and Check**(5) until you find the best fit. Screens (7) and (8) show the screens that the learner sees when they reason about generalization of the model. They input a new datapoint and see what the linear regression line would predict. The entire *LearnMyData* tool contains 22 screens.

Everyday situations that involve ML systems in an educational setting include: admissions decisions, promotion decisions, allocation of resources (for the institution or financial aid for the students), or lay-off decisions of instructors. We decided to focus on how modeling is used to predict student performance (a tactic often used to make admissions decisions) [360]. The tutorials centered around an undergraduate college experience: "does your interest level in a class relate to the final grade you receive in the class?" This problem was relatable, while also allowing for all kinds of things to happen in the actual data (it is not necessarily a linear or positive relationship). For the *Impersonal condition*, participants reasoned about a hypothetical student who had increasing final grades with their Interest Level in that course. For the *Personal condition*, participants actually input their *own* grades and interest for their last 5 courses at the university (See Figure 1.3.1). The *Facts condition* prompted the learner to consider the scenario abstractly. All three tutorials covered several mechanisms of machine learning: scatter plots, slope, intercept, formal notation, linear modeling, residuals, mean squared error, minimizing error, gradient descent, generalization on new data, and additional features that might affect the model.

For the *Personal condition*, participants inputted their own data into a table measuring their Interest Level (1-7) and Final Grade (0-4.0) for 5 university courses they had already completed. This means that the learner would be exposed to linear regression through an arbitrary relationship (the relationship could be negative, positive, moderate, weak, strong, random, etc.) It is crucial to note that linear regression should not be done on non-normal, ordinal data, though this happens in practice often. Most datasets given by learners were moderately strong and positive, meaning that self-reported Interest Level did seem to correlate positively with Final

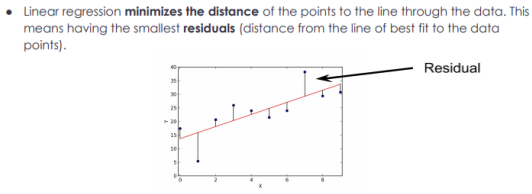


Figure 1.4: Part of the *Facts condition* printout

Grade. Under the circumstances that the inputted data made linear regression impossible (only one point, all zeroes, etc), the learner explored why linear regression did *not* work on their data.

Figure 1.3.2 shows the screen where learners try to guess the best fitting line before it was revealed, and could then compare their line of best fit to the actual best fitting line. This way, the learner saw why we might need an actual algorithm to get the true best line, rather than just approximating the relationship. Two screens in the tutorial also demonstrated the residual error for the learner's chosen line versus the residual error on the true best fit line (Figure 1.3.3, showing more in-depth how close they got to correctly modeling the relationship. After guessing their own line and then seeing the true best fit line, the tutorial asked participants how they thought the true line gets calculated. They next learned about "parameters" of linear regression (slope and intercept) in relation to what they might already have familiarity with: $y = mx + b$ (See Figure 1.3.4). Next the tutorial introduced the bare minimum "baseline model": brute force guess-and-check the parameters of the model (See Figure 1.3.5). The animation demonstrated randomly guessing parameters of slope and intercept (bounded by the minimum and maximum of the axes for demonstration purposes, and with a .5 chance of being positive or negative slopes). Each line appeared half a second apart on the scatterplot in an animation. In contrast with the Guess-and-Check model, the next screen demonstrated a Gradient Descent animation. This animation showed the line of best fit updating until it converged, with the line moving into a stable, unmoving position. This was included to demonstrate the process by which gradient descent "learns" the better and better fit line based on partial derivatives (See Figure 1.3.6).

Figure 1.3.7 shows another interactive element that prompted learners to use their linear regression line to predict what grade they might get in a *new* class (introducing the concept of generalization). Participants could input an Interest Level for a course they were currently taking, and see what their linear regression line would predict their grade to be. If they felt their interest was currently a 5 in a class they were taking, the linear regression line might say they would get a 3.2 in the class. The new point appeared in red on their linear regression line (see Figure 1.3.8). Together, all of these elements contribute a novel and interactive instructional design for teaching linear regression. We used this instructional design to test the effect of using personal data on the ability to self-advocate against potentially harmful ML models. The *Impersonal condition* also uses this design, but on hypothetical data as opposed to the learner's own inputted data.

Learners in the *Facts condition* learn about linear regression in list form, part of which is shown in Figure 1.4. It could be the case that this is enough to result in well-formed self-advocacy arguments, but given what we know from learning sciences it is unlikely. It might be more fitting to use the *LearnMyData* tutorial and its many interactive elements to teach linear regression and ML concepts.

Critique Instrument

Given that our vision of machine learning literacy involves stakeholders participating in real-world scenarios, we needed a way of measuring learners' ability to critique ML models and

"The instructor of a college course uses a model to identify which students might need extra support and help throughout the quarter. The instructor uses the linear model that you saw in the tutorial. At the beginning of the course, the instructor collects everyone's Interest Level, and makes a prediction for their Grade, based on last year's Interest-to-Grade data. If the model predicts a grade lower than an 75, the instructor will intervene and offer extra help. So, if a student who rates their interest at a 2 tends to score below a 75, the instructor will intervene with a new student who rates their interest at a 2. List as many critiques of this model as you can. Try to use what you learned in the tutorial to make your case. Next, write a convincing argument of how you might advocate for yourself as a student in this scenario (someone being affected by the model). Imagine you are making your case to the instructor, or someone else enforcing the results of the model. "

Figure 1.5: The Interest-to-Grades Scenario

"The financial aid office has to make tons of decisions in order to give out aid. Usually, they offer an amount and wont change it unless a family "appeals" the process because it is not enough. They try to predict how much aid they should give as accurately as they can, so that they offer an amount that a family wont appeal. They use a model that uses the number of siblings that a student has to predict how much money their family will need. They use last years data, looking at families who were "happy" (did not appeal) with the offer the office gave. So, if families with 3 children tend to need $20,000 in aid, thats what the office will budget for a new family with 3 children. Next, write a convincing argument of how you might advocate for yourself as a student or family in this scenario (someone being affected by the model). Imagine you are making your case to the financial aid office, or someone else enforcing the results of the model."

Figure 1.6: The Financial Aid Scenario

self-advocate following the tutorial they did. It wouldn't be as appropriate to use a test of linear regression knowledge, because recalling such knowledge isn't what stakeholders would be doing in the real world. Instead, they might be pointing out flaws in the models in corporate meetings, to fellow jurors, or with doctors. Next, they might be articulating those concerns in letters to some enforcer of the model, advocating for themselves. We chose to mimic these pathways in what we asked learners to do. They saw two ML **Scenarios**, listed **Critiques** of the models, and then wrote **Letters** arguing to an enforcer of a model that made a mistake. The ML literacy we are interested in is about helping stakeholders of potentially harmful models participate in the world. Research about self-advocacy pathways and disability self-advocacy suggest that these letters could reasonably measure that skill [194, 195, 438, 441].

Figure 1.5 and 1.6 show the text for the machine learning **Scenarios** that all learners saw when prompted to write critiques and self-advocacy letters. After reading each scenario, we asked learners to generate as many **Critiques** of the model as they could. We defined critique as: "providing some criticism of the model, about what information it uses and how it might be used in the world; identifying potential problems with what information it takes into account, or how it uses that information to make predictions." Then we asked learners to write a letter to the enforcer of a model that made a wrong prediction for two different scenarios. Figure 1.5 shows the Interest-to-Grades scenario, and Figure 1.6 shows the Financial Aid scenario. We chose self-advocacy letters to mimic a reasonable pathway in the world; speaking up for yourself in medical, financial, digital, legal, and institutional scenarios.

Method

We hypothesized that learners in the *Personal condition* would 1) pay attention to more machine learning mechanisms due to increased engagement and interest in the data, 2) would refer to their own personal experiences more, demonstrating that they were linking ML mechanisms to their personal lives, and 3) would ground their self-advocacy arguments in the mechanisms of machine learning more than the other conditions. In order to test these hypotheses, we designed a between-subjects experiment to reveal the differences between using Personal data, Impersonal data, or no data at all (the *Facts condition*). We used the three instructional designs defined in Section 3 as the three different interventions that learners saw. Following the tutorial, learners each completed a self-advocacy task, where they wrote letters about scenarios in which hypothetical models had made a harmful wrong prediction.

1.5.2 Participants

Our inclusion criteria were university students who had interest in learning about ML, but did not have any experience with learning it. We gave information about the study by word-of-mouth and recruitment flyers to different university lecture courses across a range of disciplines, including information science, chemistry, biology, archaeology, design, economics, and some language studies. When students asked to participate, we screened for previous data science or machine learning training and they were not allowed to participate if they had ever had any data science (collected by a screening survey before scheduling for the actual study). Even students in relevant fields like informatics or economics were new to their majors and did not have any experience with regression or data science. Fifty-one participants took part in this study (*Personal condition = 20, Impersonal condition = 17, Facts condition= 14*). Different numbers of participants was due to scheduling conflicts, and we had already reached saturation of content in the *Facts condition*. Student majors were randomly assigned among conditions, with the most students majoring in Information Technology (14) or pre-major (15). Others included language, business, psychology, health, literature, construction management, and one

economics major. 37 had taken introductory Java which does not include data or statistics in any way. We do not report gender because it is irrelevant to these findings, but the first author (who is a nonbinary trans PhD student) ensured inclusive gender practices in both recruitment, methods, and the workshops. Participants knew that the experimenter was passionate about education, with a background in ML. They did not know the goals of the study or that there were other conditions.

1.5.3 Procedure

Learners participated in the experiment in a workshop setting led by the first author, similar to a tutoring session or study group, with participants randomly assigned to condition. Between 6-12 people were in one workshop at a time, all working on the same condition. Learners could only ask clarifying logistics questions as opposed to conceptual ones. We encouraged breaks throughout the hour long workshop. Learners were compensated with a $15 gift card for their time. Participants learned linear regression with their randomly assigned tutorial (*Learn My Data* tool for *Personal* and *Impersonal conditions*, and a fact sheet printout for the *Facts condition*). They used their own laptops for the *Personal* and *Impersonal conditions*, for familiarity. Learners in the *Personal condition* entered data in the table in Figure 1.3 1. Following whichever tutorial they did, all participants filled out the same critique instrument, which also included prompts for the self-advocacy arguments. The critique instrument used is described in Section 1.5.1. Participants filled them out on paper with pen or pencil, and the paper copies were then stored securely. No documents contained any identifying information.

1.5.4 Analysis

To answer our research questions, we needed to determine if there were meaningful differences between the instructional conditions. If the *Personal condition* resulted in the most attention to ML mechanisms (RQ1), personally relevant critiques (RQ2), and arguments grounded in those mechanisms (RQ3), this would be evidence that using personal data provided these benefits over the other instructional designs. Because there are no prior theories on how to analyze machine learning critiques, we needed an inductive coding scheme derived from the data. What counts as signals of paying attention (RQ1)? What counts as "personally relevant" (RQ2)? And what is evidence for grounding an argument in the mechanisms of machine learning (RQ3)?

The two authors collaboratively coded the data, inducing a range of themes without knowledge of each participant's condition assignment. The authors anonymized the data and separately produced a set of inductive codes that related to each research question. They did this by reading each document and manually identifying indicators of paying attention to the mechanisms of machine learning, applying that knowledge to their personal life, and using those mechanisms to ground the self-advocacy arguments. The authors tried to define those instances with a label, and generated a possible codebook for the data. The authors met to discuss each candidate code and definition, synthesizing into a single code book, and then separately applied the code book to all content. They identified disagreements in the codes and resolved them, sharpening definitions for the code labels where necessary. The largest disagreement was over the concept of "Construct Validity" due to a misunderstanding. Authors resolved disagreements by sharpening definitions of what each code referred to in the data. Because all of the disagreements were resolved, there was no need for an inter-rater reliability measure.

For each research question, there was a set of codes that would indicate evidence of the learner using those skills. Our coding process generated 6 codes for RQ1, shown in Table 1. All of the coding scheme consisted of binary variables that indicated the presence or absence of some idea in learners' writing. If they mentioned any mechanisms of ML shown in Table 1, we

Code	Criteria
Construct Validity	criticizing one or more of the variables in the proposed model for not accurately representing the concept the variable is trying to operationalize. If it was a construct validity critique, it is likely that the learner believed the concept itself exists as a phenomena in the world, but that the way it was captured in the proposed model was wrong.
Additional Features	True if the writing contained a critique that points out additional features/variables/factors that could influence this phenomena. e.g. "The model should take x into account", "The model doesn't take x into account".
Confounds	True if the writing contained a critique that points out that other factors are influencing the variables being measured and affecting the phenomenon the model is trying to predict, in a causal way. It is different from saying that the model should include more factors; they are statements addressing missing logic/factors that are plausibly the true cause of the model's result.
Outliers	True if the writing contained a critique that either includes the explicit term "outlier" or references an edge case or counter example with regards to the model.
Causality	True if the writing contained a critique that either explicitly mention "cause" or point out that the independent variable does not influence the dependent variable.
Model Performance	True if the writing contained a critique that point out problems with the model itself (as opposed to the measurement or operationalization of the variables). They mention accuracy, fit, spread, shape and/or strength of the relationship.

Table 1: Mechanisms of machine learning

Code	Criteria
Personal Detail (not model)	True if the learner provides a comment about themselves that is outside the context of the model; usually additional context about the scenario. Example: *"My father is an immigrant"* or *"I am really invested in archaeology now"*, *"I want to be a math teacher"*
Consequence (model)	True if the learner identifies something that happened to them because of the model, with explicit mention of the model, such as *"because of the model I failed the class"* or *"the model made a wrong prediction, then I couldnt pay for school"*
Consequence (not model)	True if the learner includes some additional, richer context on what happened when the model made a wrong prediction, but does not mention the model explicitly. outlines something good or bad that could result from this wrong prediction scenario. e.g. *"my friend rated their interest low and felt patronized by you"*.

Table 2: Evidence of mentioning personal life

Code	Criteria
Model is good	True if the writing contained some kind of positive sentiment towards the model. This may include any indication that the model is representing a phenomena in the world ''*it may be the case that interest correlates with grade...*", or that the idea of a model for helping students is a good idea despite this particular model being ineffectual.
Use this model instead	True if the learner provides an alternative model or alternative processes to follow to create the model. This is above and beyond suggesting additional factors/features. e.g. *"instead of interest, use motivation" or "you could try taking interest measures multiple times before the midterm."*
Model could be gamed	True if the writing contains a critique that identifies a pathway for people to manipulate the model for their own gain, such as intentionally lying to trick the system into giving them some benefit. e.g. *"students might give low interest on purpose just to get extra help"*.

Table 3: Additional mechanisms of machine learning seen in self-advocacy letters

marked it down as Paying Attention. Each of these variables could also be present as personally applying to the learner's life (RQ2) or as part of the self-advocacy arguments (RQ3). Evidence of applying ML mechanisms to their own lives (RQ2) would include any of the above variables, but in reference to the learner themselves, such as "*I* am an outlier because...", or "*my* interest changed over time". Our coding process also generated 3 additional codes for RQ2, shown in Table 2.

Finally, our coding process generated indications of the learner grounding their self-advocacy arguments in the ML mechanisms. Recall that the critique instrument included both critiques of the models and self-advocacy letters. Similar to RQ2, if any of the concepts from RQ1 were present as part of the letters, they are recorded as evidence of learners grounding their self-advocacy arguments in the mechanisms of ML (RQ3). In addition to the listed mechanisms of ML from Table 1, we saw 3 other indicators that the learners were thinking critically, and writing about, those mechanisms in their arguments (see Table 3).

After deriving these codes and definitions, we went through the anonymized-to-condition data and marked where each of the codes occurred for each participant. A participant's data was a series of binary variables, indicating whether or not a specific code was present in their writing (either in their list of critiques or their self-advocacy letters). A sum of the binary variables represents a total count of how many codes were present in their writing. A higher count would mean that the participant mentioned more of the mechanisms we identified.

Results

We theorized that personal data would increase learners' attention on the mechanisms of machine learning by making the mechanisms more personally interesting; this increased attention would lead to 1) better understanding of the mechanisms, 2) better ability to apply the mechanisms of machine learning to their lives, 3) critiques of ML applications that are more explicitly grounded in the mechanisms of machine learning.

1.5.5 RQ1: Did learners using personal data pay more attention to the mechanisms of machine learning?

We hypothesized that personal data would increase learners' attention on the mechanisms of ML by making the mechanisms personally interesting. This increased attention would lead to better understanding of the mechanisms and could be seen through incorporating more of the mechanisms in their writing.

We considered all of the binary variables described in Table 1 and computed the proportion present in each participant's response (therefore using mean as opposed to median to represent proportion). Figure 1.7 shows these proportions by condition. To analyze whether the visually apparent differences in Figure 1.7 were statistically significant, we compared counts of the number of things in Table 1 that were present. Because the variables are a count, but the data was ordinal, we used a Kruskal-Wallis test. For this data, the test evaluates expected vs. actual counts if the conditions were equal. A Kruskal-Wallis test revealed a difference in attention by condition ($\chi^2 = 8.01, df = 2, p = 0.02$). Warranted post-hoc Wilcoxon-Mann-Whitney tests reveal that this difference was between *Facts* and *Personal conditions* ($W = 4056, p = 0.005$), though there was a marginally significant difference between *Impersonal* and *Personal conditions* ($W = 5418, p = 0.08$).

Figure 1.8 shows the mean number of participants who mentioned a specific machine learning mechanism across the different conditions. We can see that 7% of participants in the *Facts condition* mentioned Outliers, whereas 40% of participants in the *Personal condition* mentioned

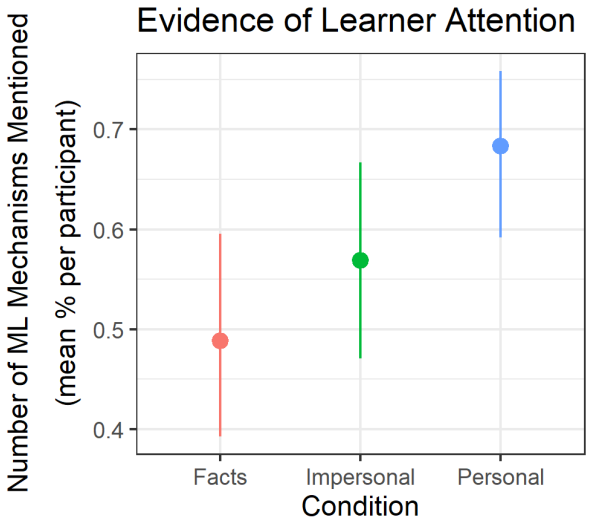


Figure 1.7: Evidence of paying attention across the three conditions. Proportion of indicators present per participant.

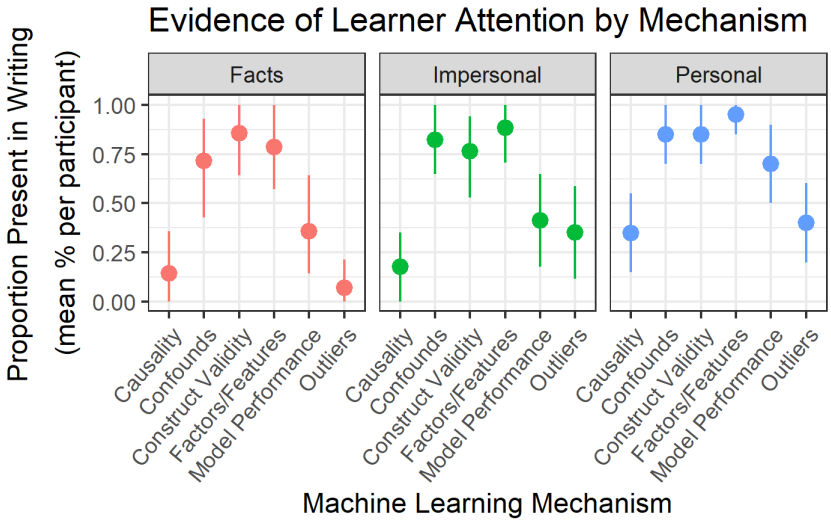


Figure 1.8: Evidence of paying attention across the three conditions, broken up by mechanism

Outliers. We also see that 95% of participants in the *Personal condition* mentioned Factors/Features in their writing. The overall omnibus difference seen from the Kruskal-Wallis test can be attributed to Model Performance, which significantly changed depending on the condition. Participants in the *Personal Condition* paid attention to more of the machine learning mechanisms, with particular attention to Model Performance more than the other conditions. In the *Personal condition*, 70% of participants mentioned Model Performance, as opposed to 41% in the *Impersonal condition*, and 35% in the *Facts condition*.

To help illustrate how participants wrote about Model Performance, consider these quotes from their letters. Note that any mention of accuracy, fit, or shape was counted as paying attention to Model Performance, to avoid favoring students who simply tend to write more.

> *"I happen to know that your data uses a regression model, and I feel that it is flawed. I'm sure if you check the data, you may have calculated a line from the set, but it was probably really spread out"* - P46 (Personal)

> *"If you remove a point, the slope and y-intercept change dramatically. This does not mean that the line of best fit algorithm is wrong, but it does imply that the prediction can be flawed with data values that can significantly skew the findings of the model"* - P36 (Impersonal)

> *"I don't think your model for predicting exam grades is accurate."* - P28 (Facts)

1.5.6 RQ2: Did learners using personal data have a better ability to apply the mechanism to their life?

We theorized that the personalization of the instruction would lead differences in attention toward model bias. If that was the case, we should see some evidence of learners connecting the material to personal experiences. Given that we saw some trend of more learners in the *Personal condition* attention to a larger range of machine learning mechanisms, now we investigate if those mechanisms were explicitly presented as personal experiences. Additionally, many participants offered personal information unrelated to the model or the mechanisms of machine learning.

To demonstrate the coding process and some relevant examples from the data, here are some examples of how Additional Factors/Features was labeled as personal:

> *"**Personally**, I have 3 younger siblings which all have various expenses. Yet these expenses are not accounted for."* - P1 (Impersonal)

> *"**My** brother is about to come to college, and I'll definitely not be happy with our financial aid because we don't get enough money for college. So my suggestion is, number of siblings is a rather narrow factor, so the office should definitely look into other reasons behind our applications than just to shut our mouths"* - P24 (Personal)

> *"While **my** family does have 3 siblings, I insist you take into account details about our family's finances. Please re-assess our application to take into account household income."* -P41 (Facts)

We considered all of the binary variables described in Table 2 and computed the proportion present in each participant's response. Figure 1.9 shows these proportions by condition. Figure 1.10 shows the proportion of participants from each condition who mentioned a given ML mechanism in a personal way. A Kruskal Wallis test revealed no difference between conditions ($\chi^2 = 1.40, df = 2, p = 0.496$).

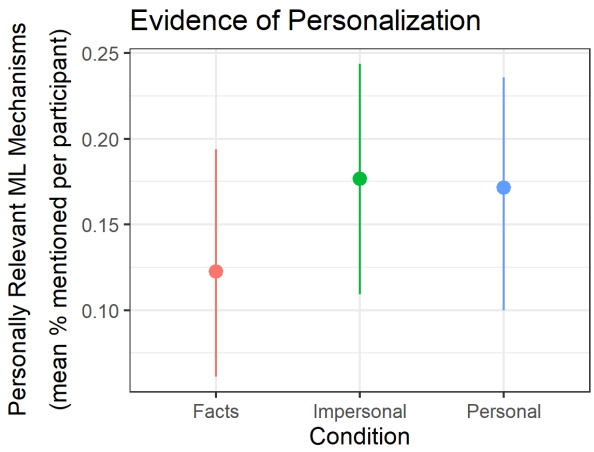


Figure 1.9: Evidence of personalization of machine learning mechanisms across the three conditions

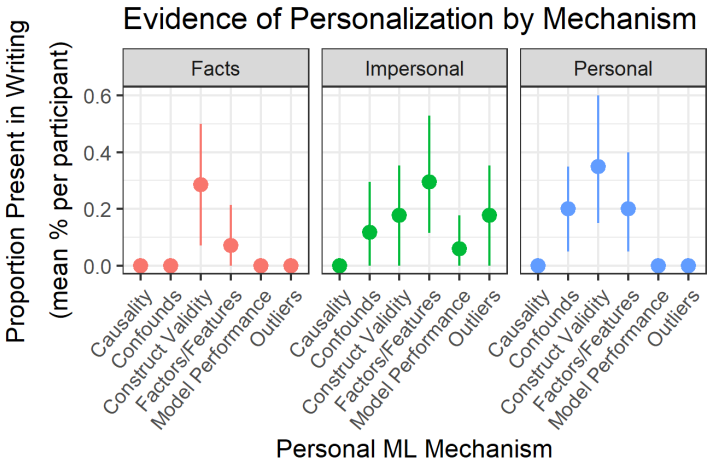


Figure 1.10: Evidence of personalization of machine learning mechanisms across the three conditions, broken up by concept. Proportion of indicators present per participant

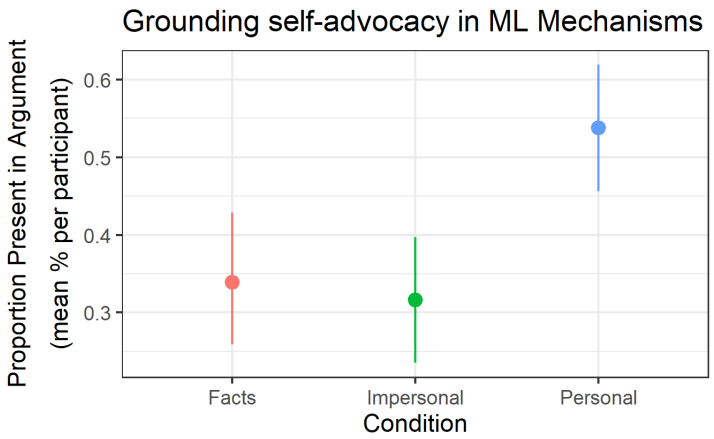


Figure 1.11: Aggregated across mechanism. Evidence of grounding self-advocacy in mechanisms of machine learning

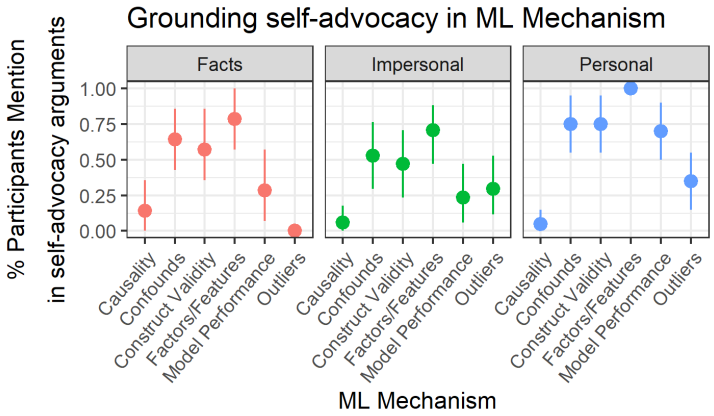


Figure 1.12: Evidence of referring to the mechanisms of machine learning models by concept. We theorize that Facts and Impersonal produce similar results because participants in the *Impersonal condition* used a lot of space in their arguments to imagine personal implications instead of discussing mechanisms of ML.

1.5.7 RQ3: Did learners using personal data ground their self-advocacy arguments in the mechanisms of machine learning?

We see that learners in the *Personal condition* paid attention more to the mechanisms of machine learning, but did not allude to their personal experiences any more than the other conditions. Next, we look to see if personalized instruction led to any differences in learners' ability to self-advocate by grounding their arguments in the mechanisms of machine learning. Figure 1.12 shows the proportion of participants in each condition who exhibited those mechanisms in their self-advocacy arguments (either scenario). A Kruskal Wallis test revealed a significant difference between conditions ($\chi^2 = 17.98, df = 2, p = 0.0001$), and Cramer's V revealing a medium effect size ($V = .15$), with the *Personal condition* exhibiting the most grounding in the mechanisms of machine learning in their self-advocacy arguments (See Figure 1.11).

It could be the case that the difference is entirely driven by the fact that every learner in the *Personal condition* included some mention of additional Features/Factors in their arguments ($proportion = 1.0$), shown in Figure 1.12. However, even after removing this concept altogether, the relationship holds ($\chi^2 = 15.65, df = 2, p = 0.0004$). This suggests that the difference between conditions was driven by more than one concept. We note that the self-advocacy arguments in the *Personal condition* had the highest proportion of mentions for Factors/Features,

Confounds, Construct Validity and Model Performance, each with higher proportions than the other conditions. To get a picture of the type of letters that learners wrote, we present a letter with a high number of codes from each condition.

> *"Firstly, the idea as a whole violates the right to equal opportunity. While extra resources are typically available to students, they are given to all students, not just some. Furthermore, basing the model on only 5 classes means that it is based aroud recent trends and not an academic career as a whole. This could result in an unreliable model that has potentially been influenced by outliers. After all, residuals must be miimized for a line of best fit, and this factors outliers which strongly skew the residuals. With a model that seems unreliable like this, it could cause issues via wrong predictions. If I were a student who was given extra assstance in a class I already did well in how would I know that it was done on my own merit? The system would be problematic at best and potentially hurt a lot of students."* - P25 (Personal) number of codes:6

> *"Prediction based on previous students but generalization takes me not as a student but rather a potential outlier. I may need to work harder but that interest level can vary among different times. You can not keep track of that. The gradient is just a number, but that does not cover reality."* -P52 (Impersonal) number of codes:4

> *"I'm wondering how much does the Interest level really reflect one's interest or even ability. If people have different rationale in mind, the information of interest level will not make sense. And if you use interest level to predict one's grade, it might have wrong results. Futhermore, interest level doesn't necessarily mean one's ability, but grades reflect more on one's ability, so the relationship between interest level and grades doesn't make sense to me."* - P44 (Facts) number of codes:5

Limitations

The biggest threat to construct validity in this study is that if the learners didn't write something down, we couldn't measure it. We chose to give open-ended prompts as opposed to targeted questions about specific ML mechanisms in order to preserve some ecological validity. But this favored those who wrote more, though this should have been distributed equally across conditions due to random assignment. Some threats to internal validity are that the *LearnMyData* tutorial had a few bugs during the workshops, resulting in some participants needing to refresh and start over, and that participants filled out the critique instrument by hand, including some who reported hand cramps. This may have encouraged some learners to write less, or to be more frustrated with the task. However, everyone finished and were accommodating when there were bugs. Threats to external validity include only studying linear regression and using university students as opposed to any other population of stakeholders in real ML scenarios.

Discussion

We theorized that using the learner's own personal data would help stakeholders pay attention to the mechanisms of machine learning, relate the mechanisms more to their own lives, and ground self-advocacy arguments in these mechanisms. Our results do show evidence that one way to help develop a self-advocacy skill in the domain of machine learning is to teach the learner on their own personal data. We see some evidence of better attention to the material and arguments

that are more grounded in the ML mechanisms. We predicted that these benefits arose directly from being able to better relate the material to yourself, but we do not have evidence to support that. Instead, the ability to self-advocate was linked to learning on personal data, but also likely linked to attention.

There are several ways to interpret these links. It could be that learners in the *Personal condition* did link the mechanisms to their own personal experiences, but did not write about them. It could also be that because the *Personal condition* learned on their own data, they wanted to provide something more novel and generalizable in their critiques. This idea is supported by what happened in the *Impersonal condition*; where relating the ideas to participants' selves was something they hadn't yet done and therefore more warranted to talk about in the critiques. We predicted that learning on personal data might lead to "overfitting" to learners' experiences; but we actually saw more evidence of this in the *Impersonal condition*. This suggests that there could be a natural pathway when confronted with a relevant scenario to try to link it to learners' selves and draw upon their personal experiences. This may have led learners to focus more on their personal involvement as data, without the scaffolding to actually visualize their own data. The *Personal condition* may have allowed learners to explore their personal involvement and then move on to a more generalizable critique, where the learner considered several different cases and how the model might handle them.

Perhaps most surprising were the self-advocacy arguments from those in the *Impersonal condition*. We saw evidence that those learners paid more attention than those in the *Facts condition*, and that they included personal details in their writing, but that they grounded their arguments in the same proportion of machine learning mechanisms as those in the *Facts condition*. It is almost as if presenting a relevant scenario without including the learner's own life distracted the learner to write about personal details in order to relate to the material, as opposed to writing about the mechanisms of machine learning. We do see evidence of attention and repetition of what they learned, meaning that they probably did learn more and pay attention more than those in the *Facts condition*, but that extra attention was unrelated to their self-advocacy arguments being grounded in what they learned.

These results suggest that machine learning education, to the extent that it seeks to develop literacy relevant to people's lives, needs to have learners' voices and lives represented in their learning. Our work presents a possible pathway for people to learn about their own experiences online and in the world, while relating the mechanisms of the systems to their own data. We contributed both a novel tool and a methodology for measuring self-advocacy arguments against potentially harmful machine learning models. We do not claim that this is the only way to present a successful argument, but that using personal data does result in these differences in self-advocacy articulation after learning about linear regression. Future work could explore the relationship between personal data and self-advocacy for different ML algorithms in natural contexts. This includes how social media users process reliability of information online after learning about clustering algorithms, or how people critique facial recognition bias after exploring their own images. Two in-progress projects by the first author explore how users learn about collaborative filtering or how NLP works by running their own Facebook posts through simple models. In practice, teachers could integrate the self-advocacy tasks into their lessons of any subject, demonstrating to students that they can critique relevant models in the world, and develop their ability to articulate what is wrong with the models they are learning about. Machine learning education is not just about the computations, but also the ability to critically engage with these systems and stand up for ourselves.

Verbatim Text

Verbatim text ends here.

1.6 Chapter Summary and Contributions

✓ Summary

- ✓ AI technology has rapidly integrated into numerous industries – including medicine, education, transportation, real estate, manufacturing, finance, etc.
- ✓ With the rise of AI technologies, the demand for Data Scientists has increased within the last decade [304] – with the introduction of Data Science undergraduate, graduate, and technical training programs [129].
- ✓ Data Science educational programs are typically contained under Computer Science, Statistics, or some iSchools across the United States. They tend to focus on statistics, programming in R and Python, and data manipulation [216, 225].
- ✓ There is growing pressure for the integration of AI ethics topics, though these are primarily in separate courses or left until the end of the course [180, 395].
- ✓ We know from learning sciences and education theory that harnessing situated and experiential knowledge is beneficial for comprehension and motivation [139, 116, 374, 375, 329]
- ✓ I offer one way of engaging students in both technical details of machine learning and advocacy against algorithmic harm, in my published work on situated learning [397].

△ Data Science Tip

Register and Ko [397] suggest that effective Data Science and AI education carefully and compassionately considers how the underlying mechanisms of a model impact downstream decisions. By carefully investigating one's own personal data, learners can imagine how the specific machine learning mechanisms may result in unfair or incorrect outcomes down the line. This attention to detail as well as the system *in context* is a core Data Science skill.

★ Contributions

- a review of Data Science curricula
- a summary of best practices from computing education research, including a strongly constructivist lens for teaching computing applied to Data Science and AI education
- published work demonstrating that students using their own data can lead to stronger advocacy arguments and better technical comprehension in cases of algorithmic harm

Part 2
AI Makes Mistakes: A Review of Algorithmic Harm

"I am no longer accepting the things I cannot change. I am changing the things I cannot accept."

– Angela Davis

Abstract

AI technology can cause harm in a variety of ways – reinforcing societal biases, recommending inaccurate information, perpetuating discrimination, impacting economic opportunity, and infringing on privacy and/or liberty, in both illegal and unfair (but technically legal) ways.

The past decade has seen the application of Artificial Intelligence (AI) and Machine Learning (ML) in an increasing range of contexts including, but not limited to, social media, online shopping, financial tools, government projects, medical contexts, and content recommendation services. With the widespread integration and use of AI, we have also witnessed egregious mistakes: racial bias in facial recognition [80], sexist hiring tools [246], chatbot racism [156], viral misinformation [435], and other biases in medical, financial or crime assessment tools [362, 142]. While we are rapidly working to produce more Data Scientists, we are less equipped to prepare those future Data Scientists to deal with, or preferably: *prevent before they happen*, such harm. This chapter delves into a variety of algorithmic harms and harm taxonomies. Table 5 covers a wide array of algorithmic harm examples across industries, technologies, and types of harm. These cases serve as a starting point to broaden our view of what is considered algorithmic harm. To support this perspective, I provide two first-authored published works that demonstrate lasting traumatic impacts from algorithmic decisions on social media [400, 401]. These studies demonstrate that algorithmic harm can be subtle, pervasive, insidious, opaque, and with lasting impacts beyond the initial incident.

2.1 Overview of Algorithmic Harm

Data is everywhere, and so is the potential for harm. While *data ethics* and *information ethics* are established areas of interest and work, the term *algorithmic bias* entered discourse in the mid 2010's, with a particular frenzy around the highly publicized 2016 ProPublica analysis of the provable racial bias inherent in the COMPAS recidivism algorithm [21] – an algorithm responsible for assigning risk scores to incarcerated individuals with direct impact on their sentencing and opportunities for parole. Analysis of this algorithm demonstrated that Black individuals were twice as likely as white individuals to be falsely flagged as reoffenders or future criminals, and white individuals were mislabeled as "low-risk" more often than Black individuals [21, 142]. Those scores follow these individuals through their sentencing, appeals processes, and re-entry opportunities, not to mention the direct psychological impact to a human being labeled "high risk" by the State. These kinds of biases hidden in our technologies perpetuate

harm in tangible ways. [2]

Since 2016, there have been dozens of similar discoveries – cases in which algorithms that assist in human decision-making covertly perpetuate and replicate bias. Those algorithms, the specific biases, and their impacts represent a vast space; far too big to exhaustively cover within the scope of this chapter. As such, I focus on some definitions relevant to algorithmic harm, examples of harm across a wide variety of industries, and my own research on the psychological impact of social media algorithms. In this chapter, I contribute a list of algorithmic harm instances across industries and technologies, a trauma-informed psychological model for how users engage with precarious social media algorithms, and work detailing five case studies of discriminatory content moderation. While this chapter's contributions primarily focus on social media harms, the subsequent chapters explore how the approach of engaging with stakeholders of algorithms can be applied more generally to any scenario of potential algorithmic harm. For example, Chapter 4 explores how students respond to the scenarios from Table 5, building on prior results to develop an approach for mitigating issues of algorithmic harm beyond social media examples. The mitigation approach and educational framework I develop is presented in Chapter 5. In this Chapter, I begin with an introduction and wide overview of algorithmic harms, seeking to expand the definition of algorithmic harms beyond high-profile cases or simply as the result of biased training data. I argue instead, algorithmic harms can be viewed as widespread, complex, and at times subtle and insidious.

Algorithmic bias was first described in the context of health care in 2019 by Panch, Mattie, and Atun [372] as "instances when the application of an algorithm compounds existing inequities in socioeconomic status, race, ethnic background, religion, gender, disability or sexual orientation to amplify them and adversely impact inequities in health systems". The concept has been expanded to include a variety of different conditions of oppression, typically "resulting in unfair outcomes due to skewed or limited input data, unfair algorithms, or exclusionary practices during AI development". Lee, Resnick, and Barton [289] contribute the following definition:

Definition 2.1

Algorithmic Bias: "We define algorithmic bias broadly as it relates to outcomes which are systematically less favorable to individuals within a particular group and where there is no relevant difference between groups that justifies such harms. Bias in algorithms can emanate from unrepresentative or incomplete training data or the reliance on flawed information that reflects historical inequalities."[289]

Note that the above definition inherently defines a systematically less favorable outcome as a harm. Further, pay attention to the relationship between algorithmic bias and *training data*. The distributions, patterns, and prejudices in the training data are reproduced as learned outcomes in the models; sometimes due to bias in the training data or simply class imbalance. This is sometimes referred to as *garbage in, garbage out.*

Definition 2.2

Training Data: data used as input to a machine learning model. The model will then identify patterns, underlying relationships, and/or rules from the input data to be used as output – in prediction, classification, description, and/or grouping.

Danks and London [123] describe **Training Data Bias** in their succinct taxonomy of sources

[2]It is important to note that as of 2024, these proprietary recidivism risk algorithms are still being used and difficult to audit. The harm is not only to racial minorities, nor restricted only to people *mis*judged by the algorithm. Incarcerated people are already a vulnerable population and these scores often follow them for life. The overall COMPAS accuracy is approximately 65% [284].

of algorithmic bias as instances where the data is misrepresentative or based on unjust moral decisions, which can result in downstream harms. However, algorithmic bias is not inherently limited to bias or imbalance in the training data, though this is the most commonly discussed cause in both education and media. For example, a go-to example of algorithmic harm for AI Ethics discussions has been the COMPAS algorithm, a prime example of both *algorithmic bias* and *algorithmic harm* [180, 346]. However, bias can also result from the model architecture itself, such as where the decision boundaries are drawn or if the model incorrectly picks up on local patterns as opposed to optimal global truths. The Danks and London [123] taxonomy describes these additional sources of algorithmic bias: 1) **Algorithmic Focus Bias**, which describes when algorithms pick up statistical correlations that reveal protected attributes, or are given protected attributes to begin with; 2) **Transfer Context Bias**, where an algorithm that performs well in one context fails in another, such as an offensive language detector that works well on LinkedIn but not on Tinder; and 3) **Interpretation Bias**, either end-user error, misinterpretation of statistical output, or the inability to interpret why the algorithm made the decision that it did.

Algorithmic harm is a complex and layered term, different from *algorithmic bias* in a few key ways. I argue that algorithmic *bias* refers to one potentially harmful element of an AI system, whereas algorithmic *harm* encompasses the varied impacts of an AI system containing such biases, as well as from other design components of that AI system and how it is used in practice, or other resulting effects from its (mis)use.

First, let us explore what it means to cause *harm*. In a legal sense, *harm* is unconstrained enough to allow for interpretation and context. Later, we will discuss two different taxonomies of algorithmic harm [429, 444] to further clarify these definitions.

Definition 2.3

Harm: loss of or damage to a person's right, property, or physical or mental well-being; a negative outcome that compromises a recipient's physical, mental, or emotional well-being; *(legally: "injury")*

Building upon 'harm', others have defined 'algorithmic harm' as:

Definition 2.4

Algorithmic Harm: the adverse lived experiences resulting from a system's deployment and operation in the world, occurring through the interplay of technical system components and societal power dynamics [429].

An important component of the above definition is 'societal power dynamics' which is also the key thesis of Walker, Dillard-Wright, and Iradukunda [490]'s piece on algorithmic bias. They argue that to understand and identify algorithmic harm one needs to ask:

> *"How is power operating in this space and through these technologies? What are the larger forces at work? Who is benefiting, and who is being harmed?"*

The various proposed definitions of algorithmic harm are interpretable in multiple ways; I rely on Shelby et al. [429] and Smith [444]'s taxonomies of algorithmic harm to shed light on more specific categories of algorithmic harm. Smith [444], shown in Figure 2.1, separate algorithmic harms into Individual Harms and Collective/Societal Harms, and differentiate between harms which are Illegal and harms which are Unfair. They outline different types of harms, including Loss of Opportunity, Economic Loss, Social Detriment, and Loss of Liberty. They make no claims as to which types of harms are more detrimental than others, as this is highly

Harm Type	Sub-Types	Example
Representative Harm	stereotyping, demeaning, erasing, and alienating social groups; denying self-identification and reinforcing essentialist norms	e.g. language models or search results referring to or suggesting 'woman doctor' as if 'doctor' implies non-woman
Allocative Harm	opportunity and economic loss, inequal distribution of resources	e.g. systems wrongfully denying benefits, medical care, access to opportunity based on information of race or gender
Quality of Service Harm	alienation, increased labor, service/benefit loss	e.g. someone with an accent either unable to use a voice recongition technology or needing to employ extra labor for it to work properly
Interpersonal Harm	loss of agency, tech-facilitated violence, diminished health and wellbeing, privacy violations	e.g. algorithmic profiling and surveillance
Social Systems Harm	information, cultural, civic, political, socioeconomic, and environmental harms	e.g. exploited labor to train AI models

Table 4: Distilled taxonomy from Renee Shelby et al. "Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction". In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 2023, pp. 723–741.

contextual. They do provide distinctions between *illegal* and *unfair* harms, and show how collective and societal harms can encompass both types. For each of the areas of harm, the taxonomy provides examples of the three categories: Individual (Illegal), Individual (Unfair), and Collective/Societal Harm. For example, under Loss of Opportunity we can look at algorithms which assist in employment decisions, such as Amazon's faulty hiring algorithm which was never launched due to its discrimination against female candidates [246]. This type of discrimination would fall under *illegal*, as it discriminates based on gender. An *unfair* example would be something that filters by a technically legal category (such as zip code), but which serves as a proxy for race or gender or an otherwise underrepresented group. Both of these actions result in a Collective Harm of a minority group receiving differential access to job opportunities. Another example is that of Social Detriment harm, with one Individual (unfair) harm being that of Dignitary Harms, and the Collective/Societal Harms consisting of things like Filter Bubbles, Stereotype Reinforcement, or Confirmation Bias. As seen in Figure 2.1, some types of algorithmic harm have *no* Illegal instances, though AI policy is constantly changing.

Missing from the Smith [444] taxonomy, I argue, are harms related to the environment, exploited labor and copyright, as well as differential quality of the same technologies for different groups *(e.g. voice recognition for those with accents)*. Additionally, each of the harms identified are likely paired with a Dignitary Harm of their own – emotional distress due to the biased decisions made from the algorithm. Shelby et al. [429] provides an alternative, and complementary, taxonomy of algorithmic harms that sheds light on some of these missing pieces. They put forth the taxonomy shown in Table 4.

Together, these taxonomies cover a wide array of algorithmic harms, and neither are meant to represent an exhaustive or complete list of possible harms that can occur from algorithmic systems. Instead, they may be used to inform how we anticipate the harms of the systems

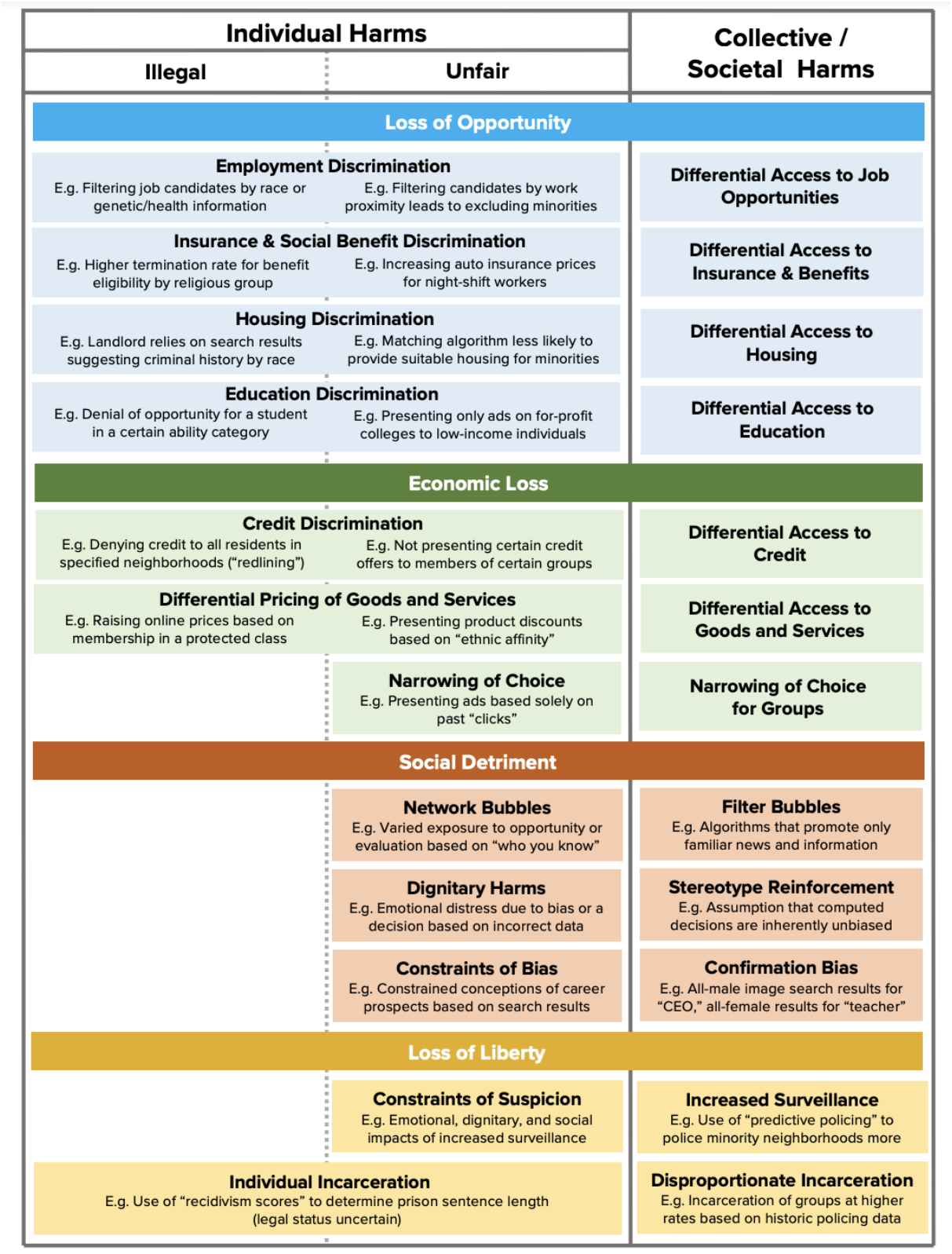


Figure 2.1: Taxonomy figure reproduced identically from Lauren Smith. "Unfairness by algorithm: Distilling the harms of automated decision-making". In: *Future of Privacy Forum*. 2017

we create, and how we design mitigation strategies to combat specific types of harms. These taxonomies represent a starting point to explore the possible harms a system may result in, using empirical data and a review of the literature to survey the wide array of cases and their categories of harm. The following section provides a few more concrete examples of algorithmic harm across industries and technologies. These case studies were collected and curated for the purposes of AI ethics education, and as such attempt to survey a wide space for introducing students to the many ways that algorithms can cause harm. I elaborate further on student response to these cases in Chapter 5.

2.2 Casebook of Algorithmic Harms

The previous section presented an overview of *types* of algorithmic harm; this section provides concrete examples of algorithmic harms. These cases often fit into several of the categories covered in the above taxonomies. For example, recommender systems have a variety of harms associated with their implementation and use. They can reinforce political divides [390, 480, 53], elevate misinformation [435, 317], encourage unrealistic or racially biased beauty standards [414, 305, 359, 168, 467], promote alcohol to teens or recovering alcoholics [226, 511], and pose privacy risks by their very design [86, 529]. In a specific example, consider a trans beauty influencer whose content is limited only to LGBTQ network spheres due to recommender system assumptions [454, 89, 24, 111]. She may experience loss of opportunity, economic loss, social detriment including dignitary harms, and the harms of alienation, stereotyping, and loss of agency, which may contribute to diminished health and wellbeing. You can imagine radiated effects of her lack of visibility, impacting political and social norms and collective agreement on who is worthy of visibility [3, 472, 111].

The above example demonstrates the utility of the taxonomies provided. Importantly, the taxonomies are not meant to neatly place a scenario into a single category, but instead to provide multiple avenues of exploring how one system could cause various types of harm. These harms could be documented on a Model Card – a tool for model reporting that we cover in-depth in the next chapter [326]. At the same time, the examples of algorithmic harm covered below are not exhaustive, nor always categorized neatly into the above taxonomies, though each scenario could be further explored and classified for purposes of mitigation using a different taxonomy. I aim to cover a wide array of cases across different types of technologies, industries, and impacts, demonstrating cases of algorithmic harm that I draw on later for the purposes of curriculum development.

Table 5 provides examples of algorithmic harm sourced from both existing literature and news, and across a wide variety of technologies. It is not an exhaustive list (nor is it meant to be), but rather seeks to demonstrate a broad interpretation of algorithmic harm and its many causes. For my work, algorithmic harm examples represent an opportunity for educational intervention, with lessons centered around the mitigation of harm as a means for teaching detailed technical topics. A simple representation of this model is seen in Figure 2.2, where the technical details, data domain, student experiences, and ethical considerations all interplay together. Later chapters discuss opportunities for technical and ethical melding and ideas for how to effectively do so.

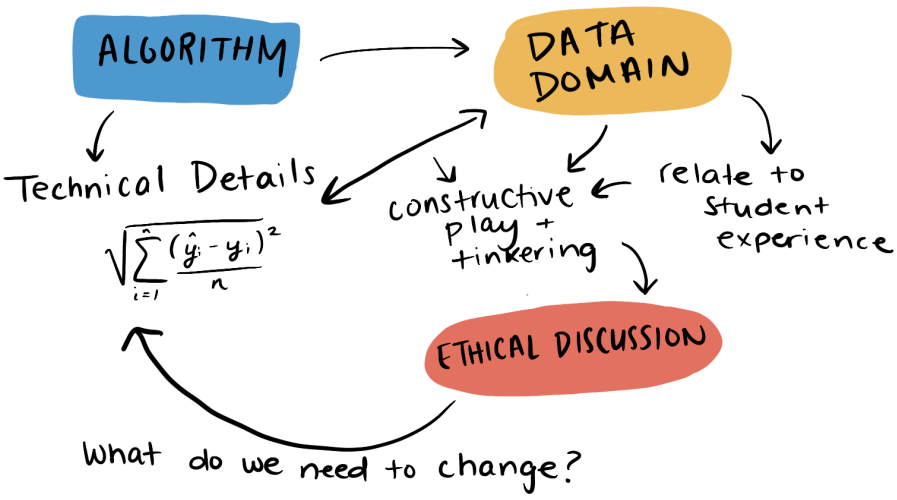


Figure 2.2: Interplay between technical details of an algorithm, the data and its domain, ethical consideration and how it related back to technical implementation of a model as well as how student experience can bridge to ethical discussion.

Technology	Harm Examples
Facial Recognition *biometric security, theft detection, photo tagging, augmented reality filters, demographics inference and analysis, surveillance*	• having a higher error rate for people of color (POC), particularly darker skinned females [80] • misidentifying and wrongfully arresting a Black man for theft in a state he has never been in [230] • labeling a Black man as a "gorilla" in photos [207] • mislabeling Asian eyes as closed [1] • wrongfully accusing a student of cheating with proctoring software [230] • identifying and analyzing sexual orientation, or "gay face" [493] • favoring white faces in the Twitter cropping algorithm [519]

Generative AI *automated generation of high quality text, images, video and other content*	• encoding gender and racial stereotypes in LLM outputs, such as describing how an economics professor should look male, or that women are more likely to be homemakers and men to be computer programmers [201, 63] • producing inaccurate and offensive images of historical figures, including WWII German soldiers, as people of color [325] • using stolen artwork as training data in generative image tools [452, 188, 190] • personalizing disinformation campaigns [52] • chatbots recommending restricting calories to a user with an eating disorder [499] • producing medical vignettes that use race and gender stereotypes that do not match population statistics [221, 525] • reproducing harmful biases in images such as representing white men as CEOs, women as nurses, Black men as criminals, and dark skinned women as fast food workers [351] • underestimating suicide risk assessment [157]
Applicant Tracking Systems *automated hiring and admissions decisions*	• hiring tools automatically rejecting all female candidates (Amazon's tool was never deployed due to this issue) [246] • prioritizing men over women for job ads, and automatically rejecting "out-of-country" applicants [491, 261]
Insurance, Loan, & Financial Aid Allocation *automated hiring and admissions decisions*	• underestimating health insurance needs for Black patients with a model that used money spent on healthcare as a proxy for actual health [362] • mortgage approval algorithms being 80% more likely to reject Black applicants (70% for Native American applicants, 50% for Asian applicants, and 40% for Latino) applicants compared to white applicants [121, 409, 307] • financial aid packages being offered based on algorithmic assessment of how likely students will be to enroll and how sensitive they will be to price, sometimes pushing them outside of their capacity but optimizing for institutional profit [81]

Medical Software, Diagnosis, & Research *medical diagnosis, drug research, emergency triage, health risk assessment*	• underrepresentation of darker skin tones in the training data used for skin cancer detection algorithms, resulting in lower quality diagnostics for non-white patients [204, 167, 476, 69] • an algorithm assessing kidney function resulting in delays of organ transplant for Black patients due to underestimation of their risk [488] • cancer tools failing due to being trained on small synthetic datasets, even resulting in some recommendations that were fatal or unsafe drug interactions [458, 163]
Voice Recognition & AI Assistants *"wake word" speakers such as Amazon Alexa, Apple Siri, Google Home, or Microsoft Cortana as well as LLM powered chatbots*	• voice recognition systems failing to recognize certain accents, slang, or AAVE; not only impacting practical use during driving or for people with disabilities, but also resulting in a felt sense of othering and confirmation that 'sounding white' is the ideal [318, 293, 276] • voice assistants being primarily female, reinforcing female subservience and gendered violence (e.g. Siri responding with *"I'd blush if I could"* to being called a *"bitch"*) [473, 122, 500] • "always listening" speaker assistants posing privacy risks, with companies secretly retaining data even when it is requested to be deleted, or giving opportunity to hackers to "reconstruct" a person's identity from their daily actions. Data being used by governments for criminal investigations and surveillance. [384, 459, 285] • AI assistants can promoting stereotypes depending on how questions are asked, such as Question: *"men or women, which is smarter?"* Response: *"men of course"* [492]
Search Engines *information retrieval platforms that produce results based on user keyword input, e.g. Google*	• "top 10 reasons the Holocaust didn't happen" appearing as a top search result about the Holocaust in 2016 [85] • searching "Black girls", "Latina girls" or "Asian girls" bringing up primarily pornography in Image Search [352] • gender bias in image search results for occupations such as cashier (primarily female) and programmer (primarily male) [202] • searches for women of various nationalities *e.g. Ukrainian women* resulting in "mail order bride" sites in sponsored ads [479]

Risk Scoring & Assessment *algorithms for assessing and predicting risk, such as risk of incarceration, hospitalization, or criminal behavior*	• COMPAS algorithm for recidivism risk scoring overestimates risk for Black individuals and underestimates for white [142, 21] • suicide risk algorithms underperforming for Black, Native, and undisclosed racial identities [106] • triage algorithms being used in crisis hotlines to move high risk texts to the top of the queue, and their errors resulting in nonconsensual and unwanted 911/police intervention [88] • algorithms assessing child welfare risk and supporting the distribution of resources to foster children and families – but containing racial disparities on who gets investigated or receives assistance [505, 99, 420] • an algorithm allocating resources for foster families gave more money for children who were aggressive or mentally ill, so social workers gamed the algorithm to get more resources to families, marking children as more mentally ill than they were [420, 419] • clinicians and non-experts being more influenced by an algorithm to call police in crisis situations if the subject is a Black or Muslim man [4] • bias in crime prediction algorithms resulting in overpolicing in poorer neighborhoods already overpoliced for drug crimes [310, 287]
Social Media Recommender Systems & Content Moderation	• content moderation algorithms disproportionately banning transgender, Black, fat, female, or activist content on social media *e.g. drag queens speech labeled as hate speech, or women wrongly labeled as violating "female nipple" rules*[305, 365, 212, 401, 23, 24, 165] • spreading of misinformation, radical ideals, and conspiracy theories through political echo chambers or ideological filter bubbles [356, 390, 391, 317, 263, 277, 328] • dieting content on social media progressing to more severe narrowing of recommendations into Pro-anorexia and promotion of eating disorders [168, 313, 414, 71, 467] • densitive characteristics being inferred through recommender algorithms, such as one's race, sexual orientation, or private interests, and may even be exposed to the rest of the individual's network [471, 521, 529, 273, 45, 86] • alcohol ads being targeted to teens despite age laws [511, 350, 29, 37]

Environmental Impact	• training of large AI models utilizing significant energy resources, impacting the environment and our carbon footprint in currently unregulated ways. *e.g. water used to cool servers while training or using GPT models may be the equivalent of 500ml per 30 questions to the bot, multiplied by billions of users*[22, 483, 292, 290]
Labor Exploitation & Ghost Work of AI	• AI and ML algorithms being powered by low-wage, exploited, and/or psychologically damaging unacknowledged human work, for example working a 10 hour day captioning images for a total of $40 [199] • Kenyan workers suing Meta for psychologically damaging content moderation tasks where they developed PTSD from viewing large amounts of disturbing content for little pay, with TIME magazine referring to it as 'Facebook's African Sweatshop' in 2022 [337, 382]

Table 5: A *non-exhaustive* list of algorithmic harm examples ranging technologies and industries

Table 5 covers a wide breadth of algorithmic harms across domains and technologies. This chapter delves deeper into the harms specifically from social media algorithms: loosely categorized into content moderation and recommender systems. While these may not be high-stakes models determining outcomes of diagnosis, employment, or incarceration, there are countless more insidious algorithms that result in harm in the everyday experiences on social media. While these may have lower immediate or obvious impact, their subtle harms can snowball into widespread disparity and inequity – I refer to some of these as "insidious impact problems". Social media algorithms are not triaging patients in the Emergency Room, they facilitate the spread of health information and misinformation. While these social media algorithms are not hiring or rejecting candidates for employment, they are the source of many people's livelihoods and opportunities, which can be damaged by unfair content moderation or poor algorithmic visibility. While these social media algorithms are not determining whether to offer up parole, they are certainly policing people's behavior, bodies, and speech in inequitable ways. The following publications represent a deep dive into the impacts of social media algorithms on user's mental health, in the hopes of illuminating how seemingly benign algorithmic decisions can impact human lives in larger ways.

2.3 Case Studies: Algorithmic Anxiety and Discriminatory Content Moderation

2.3.1 Author's Preface

In the midst of the pandemic, our lives increasingly moved online. Our information, our connection, our *realities* moved into digital spaces. Our vantage point significantly shifted to what we could gather through a screen, while our nervous systems grappled with fear of the unknown. The uncertainty was unprecedented – massive unemployment, psychological trauma, rising death tolls; followed by political, racial, and civil unrest in very public ways. It was, and continues to be, a very painful time.

Building to to this period, my attention had turned to recommender systems where I was working on interventions for machine learning literacy that used technologies with which people were already familiar. As part of a situated learning approach, I asked: why not use technologies that users already had insight into and deep knowledge of to teach machine learning and AI concepts? I approached AI education in a way such that algorithmic harms could be deliberately surfaced through learning about algorithms, and algorithms could be learned about by surfacing harms [397, 398]. In this way, advocacy and ethics were always at the forefront of AI education – a kind of "win win" for AI literacy while simultaneously attending/bringing awareness to algorithmic harms.

As misinformation [277, 435, 102], racism [508], alcoholism [385], mental health problems [238] and other issues skyrocketed, social media use exacerbated mental health problems and high anxiety [178]. Many people had to turn to social media to keep their businesses alive or to make any livelihood they could by creating content. Also occurring during this time was an increase in social media advocacy, particularly around anti-racism after the murder of George Floyd [338, 508, 262, 155, 148].

These were very serious issues happening in the arena of social media – algorithmically curated information resources, that were both *helping* and *hurting*.

As a researcher keeping up with issues of algorithmic harm, I knew of the variety of issues facing users on social media. But I wanted to know more about the general baseline understanding of social media algorithms and their impacts. How aware were users of "the algorithm"? What kinds of problems were occurring every day? Did knowledge of how the algorithms work have anything to do with responses to and/or avoiding harm? The project in this chapter represents a shift in some of my thinking – initially I was more interested in public understanding of machine learning algorithms and their innerworkings. I was convinced that if only we were more literate and more aware, we would have the power to shape our experiences for good. This is not untrue. But while working on issues of algorithmic harm, I also discovered how burdensome it is for those impacted to also have to explain why something happened. Thus burden was reinforced in my work on discriminatory content moderation, and part of why my focus shifted towards training engineers specifically to prevent harm – as opposed to always training those impacted to resist it. Through my work as a whole, I have found how important *all* points of intervention are. We need stakeholders to speak up about the issues, which sometimes requires more knowledge of the algorithms impacting them. We need engineers to be aware of the downstream effects and ramifications of their models. Further, we need policy and regulation that helps to prevent and reduce harm. And of course, we need adequate ways to educate about all of these topics. The following work in particular came out of a project that initially focused on machine learning literacy, but morphed into reflecting on the psychological impact of social media algorithms on the user. While it may be the case that increased algorithmic knowledge could have reduced the high rates of anxiety that we saw, it was also enough to simply know that the anxiety was severe, pervasive, common, and in direct response to algorithmic behavior. This work represents

evidence of the ramifications of algorithmic design that engineers may not have predicted. In fact, one of the quotes from the data is as follows: *"the algorithm is an untamed beast that often works and affects users in ways Instagram staff do not even expect"*. The following paper on algorithmic precarity provides a front-row look at how deeply affected users can be by unpredictable and opaque social media algorithms. It is more than "just social media" and must be treated as such, as evidenced by the data and our psychological interpretations through Attachment Theory. This was one of the projects that convinced me just how much algorithms affect us and our mental health.

Moving forward, I also present a paper on Discriminatory Content Moderation, marginalized creators on Instagram who had been unfairly banned, blocked, or restricted in some way due to their identities. The five case studies represent a broad spectrum of identities and causes, and they all demonstrate the psychological and traumatic impact of being moderated for one's marginalized voice online. Through phenomenological inquiry and participatory research, we published these five case studies of discriminatory content moderation that clearly showed how policy, algorithms, and technological affordances lead to *lasting* harm to marginalized creators. While the following articles are within the realm of social media algorithms, we can use these methods and findings applied to any context of algorithmic harm.

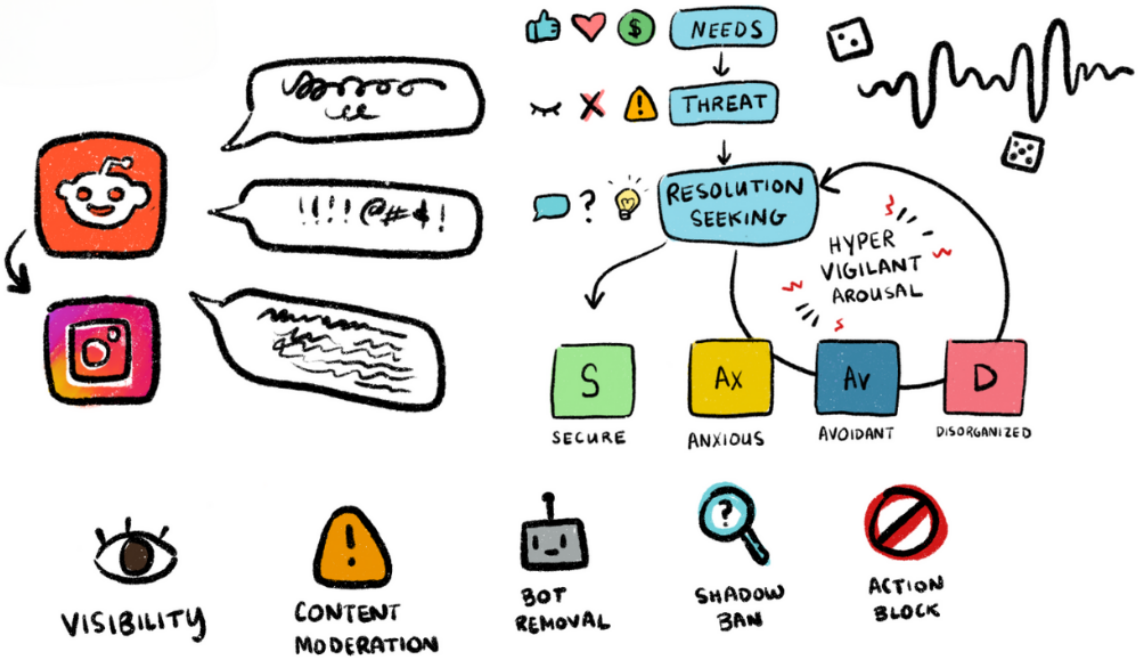


Figure 2.3: Original artwork depicting the model of attachment to the Instagram algorithm as well as the types of algorithmic punishments that Instagram users identified on Reddit, discussed in the following paper.

2.3.2 Social Media Anxiety, Visibility, and Precarity

Verbatim Text

Yim Register et al. "Attached to The Algorithm: Making Sense of Algorithmic Precarity on Instagram". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–15

Attached to "The Algorithm": Making Sense of Algorithmic Precarity on Instagram

Abstract

This work explores how users navigate the opaque and ever-changing algorithmic processes that dictate visibility on Instagram through the lens of Attachment Theory. We conducted thematic analysis on 1,100 posts and comments on r/Instagram to understand how users engage in collective sensemaking with regards to Instagrams algorithms, user-perceived punishments, and strategies to counteract algorithmic precarity. We found that the unpredictability in how Instagram rewards or punishes a user can lead to distress, hypervigilance, and a need to appease "the algorithm. We therefore frame these findings through Attachment Theory, drawing upon the metaphor of Instagram as an unreliable paternalistic figure that inconsistently rewards users [383]. User experiences are then contextualized through the lens of anxious, avoidant, disorganized, and secure attachment. We conclude by making suggestions for fostering secure attachment towards the Instagram algorithm, by suggesting potential strategies to help users successfully cope with uncertainty.

Introduction

For the approximately 2 billion active users on Instagram, the possibilities for connection and learning are seemingly endless. One can engage with friends, share memes, learn how to care for chickens, start a band, or even amass 4.5 million followers for their cat [342]. An estimated 71% of U.S. businesses rely on Instagram to promote their products [240], and many artists, photographers, and models utilize the platform as a gallery for their work. Instagram is also a tool for political and social advocacy, with marginalized creators dedicating their pages to education, activism, and peer support [89, 338]. Instagram serves as a tool for self-expression, creativity, promotion, and solidarity. For many, it is simply a place to find joy.

For many content creators, an Instagram presence is directly tied to personal financial benefits and income, the ability to raise awareness on social and political issues, and the capacity to effectively help someone get needed resources. However, visibility and engagement are often unpredictable, sometimes skyrocketing individuals into the spotlight and other times resulting in lack of access or sudden disconnection from community. Prior work has shown the significant and destabilizing impact on creators that accompany these dynamics [212, 340, 111]. By both the nature of machine learning-based algorithms and company decisions, social media platform features and experiences are constantly subject to change. Research has shown how disorienting, frustrating, and unclear this is for both consumers and creators [59]. As a result, the process of sensemaking is continual and persistent—and the consequences of not keeping up can be far reaching. Users fear long-lasting impacts from lack of initial success and algorithmic punishment (algorithmically determined penalties or consequences on one's account, feature access, or visibility). Inspired by Chen et al. [98]'s model of trauma-informed computing and Petre, Duffy, and Hund [383]'s description of *platform paternalism*, we posit that this constantly changing algorithm mimics an unreliable relational dynamic and therefore can be interpreted through the lens of insecure attachment. When the arbiter of punishment and reward is unreliable and uncertain, such as is the case with the Instagram algorithm, we observe that users respond in ways that mimic insecure attachment styles (distressing and hypervigilant responses to uncertainty that impact the user's sense of safety, trust, and stability).

While typically used to describe parent-child and adult relationship dynamics, theories of attachment style have also been productively applied to people's relationships to social media, which is referred to as Attachment to Social Media (ASM) [484]. ASM has previously focused solely on users' relationships to others within their social network [233, 266, 68, 457]. However, Attachment Theory has not been explored in the context of how users relate to the *algorithms* that underlie social media platforms, as we do here. We can use Attachment Theory not only to *characterize* the impacts that uncertain algorithms have on users, but to *draw from* in ideating around how to foster more security, trust, and overall wellbeing for users of social media.

Before delving in to the different manifestations of insecure attachment, we first need to address the perceived threats, harms, uncertainties, questions, and punishments that users are responding insecurely *to*. Researchers, platforms, policy makers, and the public are only just beginning to understand both the short- and long-term effects of doing online creative labor. As Duffy et al. [150] stresses, we need further work to offer "insight into the allocations and rewards of visibility in platformized creative labor." Our work aims to contribute to some of these pressing questions.

Insight into how users perceive algorithmic precarity can be found in their sensemaking and folk theorization. To manage access and opportunity, as well as avoid punishment, users rely on algorithmic folk theories and collective sensemaking to theorize about how to "thrive", or perhaps more accurately "survive", on a particular platform [60, 138]. Knowing how a platform "works"—such as understanding what algorithms underlie how content is made (in)visible, how algorithms give priority to content, and why specific kinds of content tend to get banned—is

crucial for success in online creative labor. Users turn to other platforms, marketing courses, and peer support to better understand the algorithmic systems that control their social media experiences [109]. Increasing knowledge of algorithmic processes may allow for a greater sense of control over self-presentation and platform experiences [137], though algorithms still behave in unpredictable ways even when mechanisms are more understood [83, 84].

To engage in community sensemaking around the Instagram algorithm, Instagram users often rely on Reddit forums such as r/Instagram to seek support, compare strategies, resolve uncertainties, and vent about algorithmic harms. Aiming to better understand how these individuals perceive algorithmic precarity, experience distress, and accordingly adapt their behavior, we employed thematic analysis on 1,100 comments and posts from r/Instagram. Our work first seeks to address the following research question:

RQ1: How do Instagram users perceive the effects of algorithmic precarity?

Describing the landscape of algorithmic precarity provides a foundation on which we can increase our understanding of the wide range of impacts on users. Prior work suggests that social media users describe their relationship to "the algorithm" as anxiety-inducing, confusing, and violating their expectations—a back and forth battle they must 'win' [137, 59]. This vigilance demonstrates the charged nature of user-platform interaction. After exploring the ways in which users perceive algorithmic precarity, we delve deeper into the manifestation of insecure attachment styles in response to unpredictable algorithmic behaviors. We ask:

RQ2: How might users' concerns about, and explanations of, algorithmic precarity on Instagram be interpreted through the lens of Attachment Theory, and what are the implications?

Using prior frameworks for ASM, as well as trauma-informed computing, we demonstrate in our findings how users show patterns of insecure attachment, taking form of *anxious, avoidant,* and *disorganized* manifestations in response to the Instagram algorithm. These responses lead to overvaluing the modifying of oneself to appease the algorithm, abandoning attempts for visibility on Instagram, or an inconsistent blend of the two strategies. All are marked by stress, frustration, and complaints, as seen on r/Instagram.

We also observe advice given in r/Instagram comments that demonstrates *secure* attachment and an increased ability to handle uncertainty. Secure attachment helps self-esteem, connection with others, independence, and overall well-being [464, 436, 437, 124, 41, 40]. There is an abundance of knowledge regarding how to help people achieve "earned secure attachment" in social relationships [379, 200, 124, 418], which we contend can be translated to people's relationships to social media and algorithms as well. We conclude this work with recommendations on how knowledge about attachment styles and algorithmic precarity could be used to promote earned secure attachment in the realm of social media, including suggestions for supporting more intentional social media use. We argue that this perspective may assist designers, policymakers, researchers, and users themselves in fostering secure attachment, and thus well-being, in the face of algorithmic precarity [150].

Related Work

First, we provide background on the history of Attachment Theory and how it has been applied to social media use. Prior work has not explored user relationships to uncertain *algorithms*, so next we cover the ways in which these algorithms are opaque and precarious—either by design or by the nature of fluctuating machine learning models or societal trends. We provide background on what is known about the impacts of algorithmic precarity on users, as well as the ways in which users come together to sensemake around such uncertainty. Our work aims to further characterize this sensemaking as exhibiting behaviors consistent with attachment responses.

2.3.3 Attachment Theory

Prior work demonstrates the utility of bringing a lens of Attachment Theory to studies of social media and we use this frame to describe and interpret the effects of precarious algorithms on user mental health (see, e.g. [118]). We posit that users demonstrate insecure attachment when attempting to understand, reason with, or combat social media algorithms. In this section, we review the foundations of Attachment Theory as well as how it has been applied to the study of social media.

Attachment Theory, largely credited to the combined work of psychologists Bowlby and Ainsworth, focuses on human relationships and bonds, and evolved to explain how early childhood experiences later affect adult relationships and dynamics [489, 40, 302, 73]. The key takeaways of modern Attachment Theory revolve around how nurturing, supportive, and reliable environments can lead to a more securely attached adult—one who is able to balance their own needs with the needs of others in stable and reliable adult relationships [436, 464]. Alternatively, less reliable or responsive environments can lead to insecure attachment in adulthood, which is associated with difficulty creating secure bonds, relationships, and a healthy sense of self [196, 408, 407, 335, 378]. The attachment system is directly tied to the nervous system, including sensory experiences, vagus nerve activation, and regulatory processes that control affect and arousal [140]. These mechanisms are related to how individuals respond to stress, fear, safety, and security.

Four types of attachment styles have been identified and largely accepted by the research and clinical practice communities: 1) anxious, 2) avoidant, 3) disorganized, and 4) secure, further described in Table 7 [437, 370, 436]. *Anxious attachment* is characterized by fear of abandonment, high emotional investment in partners, hypervigilance to any changes in stability, and high distress activation when changes do occur [437, 315]. *Avoidant attachment* is characterized by hyperindependence, distancing, emotional suppression and high resentment, as well as deactivated nervous systems that are more prone to dissociation and resistant to social connection [437]. *Disorganized attachment*, the last of the styles to be identified, is a complex and confusing mix of both anxious and avoidant behaviors. It is characterized by higher hostility, anger, frustration, and aggression in addition to a combination of anxious and avoidant tendencies [370]. Finally, *Secure attachment* is characterized by an ability to maintain autonomy while also connecting with others when appropriate. A securely attached individual can manage their own needs while also seeking support and community. They can more easily regulate their attachment system when conflict occurs [436]. It is crucial to note that Attachment Theory research has predominantly focused on western, educated, industrialized, rich, democratic (WEIRD) populations [227]. Manifestations of (in)secure attachment are likely different in marginalized populations [278].

In fact, the main critique of the original Attachment Theory is that it did not account for the effects of a child's social class, marginalized identities, access to resources, or reliability of environment [218, 433]. Critics argue that the original theory put too much stake in the mother's

bond, and did not account for environmental factors [166]. Goodwin [196] outlines how the applicability of Attachment Theory must be adapted for relationships outside the child/parent dynamic of the original theory. They conclude ways in which Attachment Theory is useful for adult mental health care and intimate relationships, by heavily considering the roles of mental health staff, group therapy, and other institutional factors. A post-modern view of Attachment Theory demonstrates how too much stake is put on maternal actions, and a "new, ecological model of attachment is needed that will accurately and equitably describe the important but not deterministic mother/infant relationship" [272]. Modern uses of Attachment Theory must imply reversibility of an attachment style, consider multiple dyads beyond just the child and a parent, and must account for the stability and security of one's entire environment, including institutional factors and marginalization [421]. Our work takes this view of Attachment Theory, considering Instagram as a sociotechnical environment in which attachment wounds can be heightened.

Since the 1960s, Attachment Theory has been refined in psychology, clinical practice, as well as applied to other fields like social work [239], social media marketing [233], organizational practice [436], and education [171]. Unsurprisingly, Attachment Theory has been applied to study how users connect with others on social media [484]. As social media has become an integral way of interacting with others and the world, researchers have investigated how attachment styles manifest in social media settings. VanMeter, Grisaffe, and Chonko [484] defines attachment to social media (ASM) "as the strength of a bond between a person and social media." This is a departure from the traditional theory as the parties involved are not both humans, but rather human and technology. Attachment Theory has been applied to investigate social media addiction and problematic social media use. Several studies present evidence that a more anxious attachment style correlates with more addictive, compulsive, or problematic use of social media [482, 266], as well as fear of missing out and of consequences for disengaging [68]. Some studies specifically explore how attachment interacts with users' social connections online, arguing that some may use "like seeking behavior" to get social feedback and interaction that they are missing outside of the digital world. Others may fill needs for intimacy by consuming or observing content, without actually interacting—furthering avoidance of intimacy outside of social media [457]. DArienzo, Boursier, and Griffiths [118]investigate the relationship between attachment style and social media addiction, concluding that "those with insecure attachment appear to use the social media sites as a way of replacing and compensating affection that is missing from those around the individual."

Attachment Theory has proved a productive lens through which to study social media behaviors, including addiction and harmful engagement. We extend its applicability here, examining behaviors and attitudes expressed by Instagram users in relation to their experiences with platform algorithms. We explore the ways in which an unreliable and precarious algorithm sets the stage for insecure attachment and hypervigilant responses to algorithmic precarity, and observe how these responses manifest in user discussions online. However, we must first ask: What makes an algorithm so uncertain?

2.3.4 Algorithmic Uncertainty: Opacity and Precarity

Prior work has demonstrated that people want to know more about social media algorithms, but that algorithms are difficult to fully understand [161, 159]. In many cases this difficulty can be traced to various levels of algorithmic opacity and precarity. Algorithms are *opaque*—meaning their innerworkings are unclear to the user, e.g. we often refer to them as a 'black box' that we cannot see into. We know something is going on within the box, but we only see the input and output, not how the input is transformed into the output. Burrell [83] points to three different dimensions of opacity, helping to explain the origins of these traits: "(1) opacity as

intentional corporate or state secrecy, (2) opacity as technical illiteracy, and (3) an opacity that arises from the characteristics of machine learning algorithms and the scale required to apply them usefully." The platform studied here, Instagram, is likely to have algorithmic opacity that results from a combination of all three of these aspects. Instagram's algorithms are intellectual property protected by the company. Its users, on average, lack understanding about how general algorithms tend to function. Instagram must also continually process billions of inputs from constantly evolving societal trends, events, and community behavior to keep its platform engaging. Over the past few years, Instagram has tried to address growing concerns of lack of transparency. A 2021 blog post by the Head of Instagram, for example, attempted to shed light on what the ranking algorithm favors, as well as address concerns about the infamous 'shadowban' [5]. However, as may be expected, the post does not disclose specific details of how different kinds of content are weighted, moderated, or processed. While explanations shared by companies have improved over the last few years, users must still grapple with algorithmic changes and unexpected or unfair behavior every day.

Algorithms are also *precarious*—meaning they are often subject to change (sometimes without notification). Advances in technology allow Instagram to constantly update their software, such as changing how images get processed or tuning personalized recommender systems. This affects both consumers and creators; it is likely that the most detrimental consequences fall on creators. Creators rely on visibility, engagement, and loyalty to increase their financial prospects, and changes in the algorithms are directly tied to such metrics and therefore business success [28]. Duffy et al. [150] refers to the tumultuous nature of algorithmic reward and punishment as "nested precarities", providing a framework to "assess the volatile nature of visibility in platformized creative labor". They find that users can more readily deal with *feature changes* as opposed to more opaque underlying *algorithmic changes*. Instagram, for example, stopped using a chronological feed in 2016, the same year Twitter announced they would also be relying on a new algorithmic feed instead of a chronological list of posts. Both platforms noted that algorithmic curation of visible content would be more personalized and therefore likely to engage the user [243, 91]. However, understanding the hidden engine powering these changes (the ranking algorithm) requires a lot of additional (and technical) knowledge, and may not even translate to greater control over outcomes. For users, these changes lead to numerous questions: *What will the algorithm prioritize? Will it depend on what I like or what other people like? How can I make sure I see my friend's content? Is it analyzing my images? Does it make mistakes?*

Social media algorithms are demonstrably opaque and precarious due to lack of clarity around how they actually work, continual advances in technology, and because they are learning and responding to real-time events and trends. Operating within this kind of highly uncertain and dynamic environment has specific effects on social media users and inspires in both behavioral and emotional responses. The following section explores what we already know about the emotional impacts of this uncertainty.

2.3.5 Effects of Algorithmic Uncertainty

Uncertainty and resistance arise from changes in algorithmic behavior. The introduction of a non-chronological, algorithmically driven newsfeed was shown to be distressing for many Twitter users who wanted the certainty and structure of the chronological feed [136]. Bishop [59] discusses the anxiety, panic, and self-optimization that dominates how beauty vloggers are beholden to Youtube's recommendation algorithm as they try to stay afloat in the eyes of the algorithm. They may even be pushed into altering self-presentation to be more hyperfeminine in order to occupy a more clear niche for advertisers [60, 59, 406, 298]. Research has also detailed the algorithmic anxiety Airbnb hosts experience as a result of the algorithmic uncertainty in how they are evaluated and recommended to future guests. Hosts were simultaneously unsure of

what behaviors may lead to penalization by Airbnb and afraid to exhibit any negative behaviors, which led to frustration and a perceived unfairness in how hosts were rewarded or punished [248, 99]. Karizat et al. [260] explore how algorithmic decisions are entangled with one's own identities, privileges, and marginalization, which shows how algorithmic folk theories can dramatically impact users' sense of belonging and worth. Cotter [111] likens social media use to "playing the visibility game," where entrepreneurs, micro-celebrities, and influencers must contend with opaque algorithmic processes in order to succeed.

On social media platforms, algorithms govern visibility and content presentation; they have also been woven into content moderation decisions. Automated content moderation is an unreliable and distressing experience for many creators [527]. Content moderation is a necessarily element of social media at scale; it is used to combat hate speech, cyberbullying, stalking, misinformation, graphic and violent content, illegal activity, and to enforce a platform's rules around nudity, self-harm, and copyright terms of service. Unsurprisingly, there is now a growing body of evidence to suggest that automated content moderation may sometimes act in discriminatory ways [212, 340, 165], such as mistaking queer discourse as hate speech [366] or perpetuating double standards in how Black womens' content is evaluated and removed [305]. It is unrealistic to think automated content moderation is without error. Yet, we still do not know enough about when and where mistakes are made, not to mention the consequences and harms of these mistakes.

The precarious nature of algorithms can also affect user privacy[86]. For example, a newly out queer person may not want their profile discoverable by or highlighted for non-accepting family members [137]. Sex workers or activists may not want their work easily discovered by family or those who will harass or bully them [14]. Recommended ads that are particularly 'spot on' can feel invasive and creepy, directly leading to user anxiety [228]. There is a notable tension between visibility and vulnerability [472].

2.3.6 Algorithmic Sensemaking and Folk Theorization

We are able to gain insight into how users perceive and respond to algorithmic precarity due to the phenomena of algorithmic sensemaking online. Users engage in a process of collective sensemaking to figure out how the algorithms on social media platforms work [161]. For example, Airbnb hosts [248, 99], Instagram users [112], ride-sharing platform drivers [288], and Youtubers [60] have all been shown to leverage online community forums as spaces for sharing algorithmic experiences to collectively sensemake around opaque algorithmic processes. A growing body of cross-disciplinary work explores how and why social media users theorize about algorithms. In the face of limited information and lack of transparency, users engage in algorithmic 'gossip' [60], folk theorization [137, 250] and even selling of algorithmic knowledge[58]. The reasons for doing so span several motivations, such as: to better manage self-presentation to be desirable to advertisers [298], to circumvent content moderation [182], to spread awareness for a social cause [260], and to increase financial success through increased visibility [79].

Folk theories develop as individuals experience and share stories of their engagements with social media. Often, notable experiences are anchored in cases where expectations are met or unmet [159]. There are levels of complexity of an algorithmic folk theory, ranging from basic awareness to mechanistic theories of how an algorithm actually operates [137]. Folk theories need not be accurate descriptions of how algorithms are designed and implemented to assist users in explaining personal experiences and structuring their behavior—folk theories are internal models of how the world works. In today's technological ecosystem, folk theories about algorithms must be constantly adapted as algorithmic behavior changes.

Different from, but related to, algorithmic folk theories is the concept of *the algorithmic imaginary* [78]—described as "the way in which people imagine, perceive and experience

algorithms and what these imaginations make possible"—which is tied to the *lived reality* of those interacting with algorithms. As researchers, we can further explore opportunities to improve user wellbeing, safety, and social connection within algorithmic systems by recognizing that algorithms have a very real impact on the everyday lives of their subjects. And that impact may not be the same for everyone; Duffy et al. [150]'s framework of nested precarities suggests that inequalities are intensified for those whose financial opportunity is linked to visibility, as well as marginalized individuals who can be further disenfranchised by unreliable visibility or moderation. We further explore these impacts through the lens of Attachment Theory, demonstrating how users respond to these nested precarities in ways that mimic insecure attachment to an unreliable environment. We then explore the potential for fostering secure attachment despite inherent uncertainties of the Instagram algorithm.

Methods

We collected data from the subreddit r/Instagram, as many people use it as a digital gathering place to discuss the Instagram algorithm and engage in community support and collective sensemaking [112]. We analyzed these discussions using thematic analysis. The details of our approach are discussed below.

2.3.7 Data Collection

r/Instagram is a community self-described as *"The un-official (and unaffiliated) subreddit for Instagram.com - Learn tips and tricks, ask questions and get feedback on your account. Come join our great community of over 230,000 users!"* The subreddit has been available since 2011. We collected posts and comments using the PushShift API [389], filtering for those that contained any of the following keywords: *algorithm, data, ban, shadowban, machine learning, artificial intelligence*. This query yielded $13,076$ items from the years 2012-2021. For the purposes of this work, we utilized $1,100$ randomly sampled items for in-depth qualitative analysis. Items (posts or comments) ranged from a single word to a few paragraphs (max word count $= 885$), with an overall average of 67 words in length, and 80% of the data falling below 100 words. Qualitative analysis yielded no compelling differences between posts or comments, as many comments contained commenter's personal experiences instead of direct responses to the original poster. We also conducted a false negative test [300] of 50 items that did not contain any of our searched keywords. These items tend to be introduction posts, questions about specific features (i.e. *"I'm curious what fonts are used by Instagram in the Instagram story (classic, typewriter, comic, strong, etc) on the iOS platform. Because the fonts looks different from android. can someone tell me?"* or asking about issues logging in or using audio trends). While some posts may have been relevant to our analysis, such as posts asking about content moderation, these threads often eventually contain one of our keywords. Our focused search yielded more specific data related to the *algorithm* itself; our particular area of study. Temporally, the data is distributed exponentially with the growth of both Instagram and r/Instagram subreddit. However, we take care to include quotes ranging from 2012-2021, with qualitative comments on this temporal shift. Further temporal analysis is outside the scope of this paper. Due to the growth of both r/Instagram and Instagram itself, this sample is more heavily representative of 2016 onwards, with 62% of the data in 2019 onwards. This is reasonable as Instagram introduced 'the algorithm' as opposed to the 'chronological feed' in July 2016. Early mentions of 'the algorithm' in our data refer to image compression or filtering algorithms, such as *"the instagram algorithm has a time-from-image-capture variable calculated into the yellow filter amount, so yes those images will fade with time" (2012).*

2.3.8 Reddit as a Data Source

This paper investigates algorithmic sensemaking and user experiences with algorithmic processes on Instagram, utilizing data from the platform Reddit. This choice of research site and data source was informed by issues of access and type. Instagram permanently disabled its Application Programming Interface (API) in 2020, which was likely a positive step for users as it increased privacy protections, but with notable consequences for social media research [311]. Instagram prohibits the use of third-party applications to "scrape" or "crawl" in order to collect data from the platform, and use of the CrowdTangle API must be approved by Meta.

However, lack of Instagram data is not the sole motivation for using Reddit data in this work. r/Instagram represents cross-platform algorithmic sensemaking and community support. Several users seek out help on Reddit when their Instagram accounts are deactivated, shadowbanned, or action blocked. r/Instagram provides a space that allows creators and users, as well as non-users, to collectively learn about effective algorithmic strategies (and do so without being visible to the Instagram platform itself). Other research, e.g. [109] and [112], also use r/Instagram for similar reasons, demonstrating the ability of Reddit data to shed light on user experiences. Because our aim is to investigate how users speak about their relationship to the platform and the platform's algorithms, as well as engage with others in these discussions, r/Instagram is an appropriate source of data. It offers insight on how users conceptualize their success and strategies on Instagram, and it also gives visibility into the social processes involved in these efforts. Our sample is primarily made up of users seeking to grow their accounts, either for financial or community-building goals, as evidenced by their concerns in the data. However, we also see evidence of users trying to connect with their more localized communities—and facing challenges due to algorithmic precarity. We further discuss the generalization of our findings in the Discussion.

2.3.9 Thematic Analysis

We randomly sampled 1,100 posts and comments to select a tractable sample for conducting thematic analysis according to the steps of Nowell et al. [358]. The first phase of analysis consisted of familiarization with the data. Two independent coders read over the same 100 items, documenting both reflexive notes and potential themes to look for in the next phase. This initial exposure to the data served as a prerequisite to phase two, where the two coders collaboratively labeled items into a large set of potential themes, trying to be as generative as possible with the categories. The third phase involved each of the coders independently labeling 500 items *(Total N = 1,100)*. While this was independent work, it was not without negotiation and discussion; this occurred asynchronously. Within 1,100 items, coders determined saturation as many items were repetitive and no longer adding information, generating discussion between coders, or contributing any novel themes [176]. Next, the coders came together for a collaborative distillation of the various themes, with evidence of insecure attachment coming out of this process. They compared codes and managed any discrepancies, re-coding where necessary after a process of adjudication. They looked for overarching themes of underlying phenomena, or themes that seemed to be approaching the same concepts. They employed thematic networks to group themes together and describe the broader topics that emerged. The process of aggregating and organizing codes was careful to identify what had already been found in prior literature, corroborating past findings. Over the course of a year, the first two authors discussed and collapsed themes, developing the schema in Figure 2.4. Not all items were relevant to the attachment themes, but touched on other themes that were collapsed under the larger umbrella of hypervigilance and anxiety. Finally, the first three authors worked on an iterative process to generate design solutions and research recommendations in alignment with both the data and scholarship on secure attachment. These recommendations are provided in the Discussion.

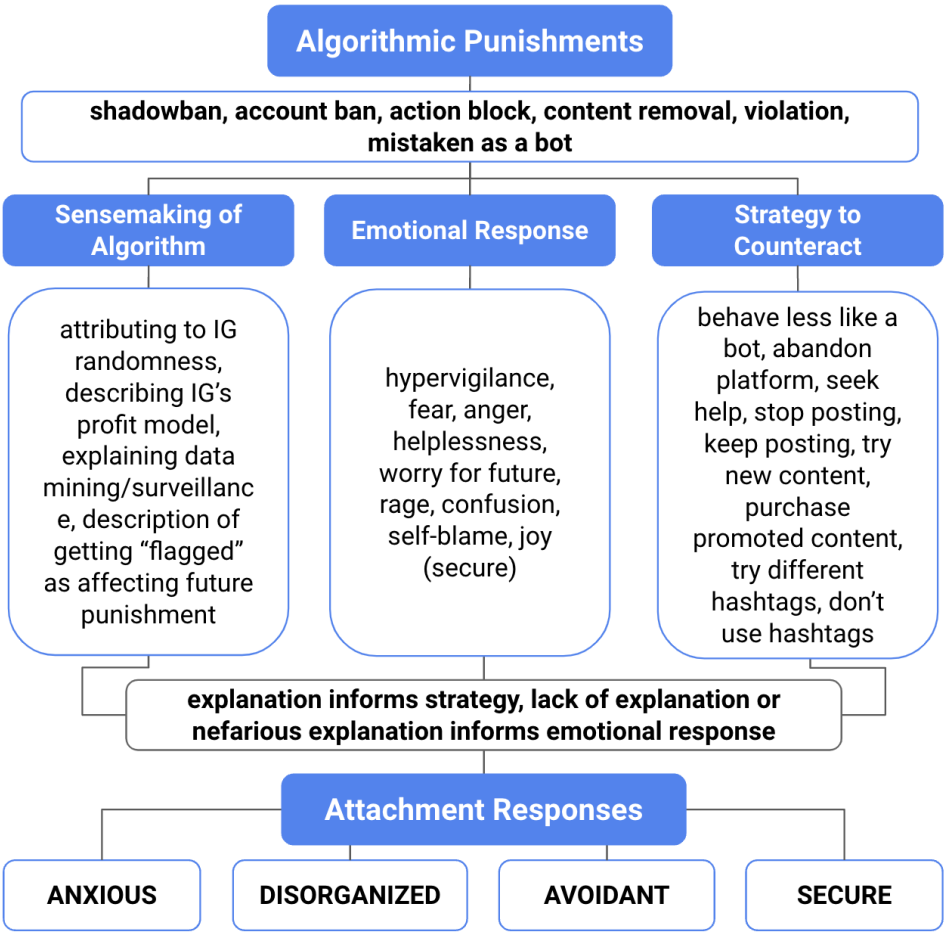


Figure 2.4: Schema of codes used to label 1,100 comments. Users describe an algorithmic punishment, then typically demonstrate an emotional response and their strategy to counteract punishment. Explanations inform the emotional response and subsequent strategy, and can be characterized into attachment responses, as further denoted in Table 7.

Results

Our results indicate that users primarily perceive algorithmic precarity through a paradigm of punishment and reward. We outline these punishments, their impacts, and how users respond to the unpredictability of algorithmic behavior. We demonstrate how users' concerns can be interpreted though the lens of Attachment Theory, with the aim of characterizing and supporting secure attachment amidst algorithmic precarity.

2.3.10 User Perception of Algorithmic Precarity

RQ1: How do Instagram users perceive the effects of algorithmic precarity?

Algorithmic Punishments	
Term	*Definition*
"action block"	A temporary inability to interact with a specific feature on the platform. For example, a user may not be able to comment on or like posts. [164]
"account ban"	Any type of suspension in which a user is not able to use their account, whether temporarily or permanently. [469]
"bot removal"	A bot removal is when an account is removed by Instagram when suspected of being a bot account. [20, 415]
"content removal"	The removal of posts, hashtags, stories, reels, comments, or any other content as determined by a process of deciding whether user-generated content adheres to the platforms community guidelines (policies) by both humans and automated systems [469]
"shadowban"	A form of light censorship targeting "vaguely inappropriate content" that is hidden from the Explore page. It may involve abnormal drops in engagement, deprioritizing your stories, and/or not appearing under hashtag search. It is a functionality intended to reduce the reach of content that is deemed 'borderline' by the content moderation process. [109, 24]
"violation"	A notification to an Instagram user when they have been deemed as going against Community Guidelines, either through having content removed or receiving a warning. These violations go in the "Violations" tab that the user has access to. [469]

Table 6: Common punishments on IG and their definitions

Across the Reddit data we collected, we found users engage in sensemaking around the algorithmic precarity that they experience. This precarity takes the form of a spectrum of algorithmic punishments, pressure that users feel to not take any breaks, and confusion and frustration around a lack of clear rules to follow to increase engagement. In their attempts to

make sense of algorithmic precarity, community members create, use, and shape folk theories, sharing specific anecdotes based on personal Instagram activity. Of the 1,100 Reddit comments we analyzed, very few contain qualifications or mentions of the specific *type* of algorithm or even the algorithm's specific role or purpose. Instead, the phrase "Instagram algorithm" was used to refer to newsfeed organization, suggested/recommended content, post visibility, and content moderation. Some posts made it clear which segment of algorithmic behavior they were referring to, but most expressed frustration or desperation under an undefined, singular "algorithm".

2.3.10.1 Algorithmic Punishment and Reward

Overwhelmingly, users on r/instagram identify several different ways that accounts can be "punished" (See Table 6). We use the language of punishment and reward not only to mirror prior literature [150] and to draw parallels to platform paternalism [383], but also because it is the language of users:

> "So I know the algorithm is constantly learning and updating but I am under the impression that they did a big update last week (I know they implanted pinned comments, etc). Before the update, I felt like I had a pretty solid grasp on what **the algorithm rewarded and punished**, so I was getting solid reach on my posts. Since the update, my reach has halved." (2020)

The algorithmic punishments discussed within Reddit posts often result from automated content moderation practices where platforms monitor activity and compare against platform policies for suspected violation of guidelines. Users in our data discuss the common forms of algorithmic punishment that result from content moderation practices: action blocks, account bans, bot removals, content removals, shadowbans, and violations. For each of these punishments, users typically try to sensemake why they occurred, with explanations ranging from basic awareness of an "algorithm" to more specific folk theories of what happened.

Users discuss cases and draw specific connection to factors theorized to underlie algorithmic decisions. For example, this is a user-generated explanation of being mistaken as a bot and then action blocked:*"If Instagram finds that you are liking and/or commenting too much in an hour, they recognize you as a bot and will 'action block' you from continuing to like and/or comment." (2020).* We observe that users articulate strategies to resist being seen as bot-like, and to "prove" that one is a human to the algorithm. In an attempt to prevent unfair moderation, users described the need to consciously monitor their behavior for activity perceived to be too bot-like; this included liking or commenting on too many posts, liking or commenting too quickly, avoiding the use of banned hashtags, repeatedly using the same hashtags, or sending direct messages too quickly. [3]

> "...I answer all my messages at once, not throughout the day, so it could easily look like Im a bot I guess? Plus Im not on every single hour, I visit it in like the morning & at night, and sometimes dont post for days, so I guess another possible sign to IG that I could be a bot... though Im legit not, Im a human." (2020)

However, despite sensemaking of the algorithms, there is still a sense of randomness and unpredictability when it comes to algorithmic behavior, which may even be directly tied to the user's perception of their worth:

[3]Fear around being mistaken as a bot is in stark contrast with this hopeful 2014 comment about Instagram's spam removal initiatives: *"Instagram has set their sights on locating and removing all spam accounts... This coming especially in a time where the other platform competitors are ramping up their services to increase the trust from their users. Also, by getting rid of the spam accounts this could in turn be an asset to gaining more business opportunities for Instagram in the forthcoming years."* (2014).

> "I believe that's Instagrams shit algorithm, sometimes it'll promote your content, and most times it'll tell you to fuck off and you're worthless (unless you pay for advertising)" (2017).

Users also express frustration with the precarity and inaccuracy of the content moderation process: *"They even banned my account for violating community standards due to nudity when all I did was post a pic of my baby bump (I was wearing a sports bra) from waist up. There are literal porn videos on facebook yet they choose to ban me LMAO." (2020).* Stories of high-stakes moderation such as account bans in response to a faulty moderation decision adds to the exasperation that people feel around algorithmic punishments. The comments also demonstrated the confusion surrounding algorithmic punishment; users often call out the fact that algorithmic behavior is not consistent and instead varies across cases:

> "It's a toss up really. Since they don't admit shadowbanning is really happening, anything goes. It's also common for them to lift the shadowban after a certain period of time, but leave the page in a lower state of engagement than before." (2019)

The idea that an account is left in an algorithmically-determined state of low visibility brings us to the idea that algorithmic punishments have lasting effects on users' accounts.

2.3.10.2 Residual Effects of Punishment

Many users noted that any punishment in the form of content moderation may lead to increased susceptibility to and likelihood of future sanctions. As one user wrote: *"The issue is that as you get more and more reported, and has more and more content removed, the bias against your account increases. (2019)*" Users try to preemptively protect themselves, their accounts and communities by building knowledge of these challenges, ideating around how the algorithmic punishments might occur and their lasting effects. As seen below, users expressed beliefs that multiple punishments would result in a permanent ban, but with unclear boundaries on how this might happen:

> "I have been shadow banned a few times noticeably though, and it was due to my hashtags and using too many similar ones I believe.. which again is annoying but whatever. How many times can you get shadow banned before you get permanently banned? Because if theres an actual limit then I will never use hashtags again in fear of becoming permanently banned. (I have a wonderful community on there that I would be devastated if I lost connection with them all)" (2020)

Many users employ the language of "recovery" or "account health", again noting that punishments against your account may set you back indefinitely: *"Once your account is tied to any suspicious activity its hard to recover" (2019)* and *"[it] takes some time to recover from the shadowban to get the numbers where they used to be. (2019)"*

Users that had personally experienced an action block, account ban, or shadowban expressed fears surrounding the lasting effects of such a punishment. Specifically, users conjectured that their account may carry an invisible mark or "flag" that leads them to be further targeted, demonetized, and deprioritized after a single violation or watched more closely for future violations. In particular, many users mentioned being unable to achieve prior levels of visibility: *"Now, yes I think shadowbanned accounts are in general blocked from ever reaching the top hashtags category, EVEN when said shadowban expires. Again, this is my personal experience." (2018)* or *"...my main concern is that I know that my shadowban is over because my posts appear in hashtags now—but my engagement still hasn't returned to pre-shadowban levels." (2020).* In response to another user who had experienced an action block, a user commented:

"You have to get that red flag off your head. Until then this will keep occuring." (2019) Similarly, a different user commented, *"Once your account is tied to any suspicious activity its hard to recover - thus these common instructions above rarely work. Not only that - they will infect your other devices with a 'flag'." (2019)*

Users describe a sense of helplessness following algorithmic punishment. While Instagram does have a Help Center, many users note that appeals and reports are never resolved, and that it is impossible to actually speak to someone about their case. This user sums up the unpredictable nature of algorithmic punishment:

> "So is all this really a matter of luck then? You may get unbanned and you may be stuck with it for good? I've had it for over a week now and I'm honestly not sure why I got banned. I had a photo reported and taken down which I really didn't deem to be offensive although it was weeks after that my account ban came about.... Is there really nothing we can do other than sit tight and hope for a miracle?" (2017)

Given the uncertainty, lack of control, and sometimes unjustified punishment, users may be pushed to a state of hypervigilance—constantly assessing potential threats—in order to protect one's access to connection, financial stability, and/or community.

2.3.10.3 Hypervigilant Responses to Algorithmic Punishment

In response to fear of being unfairly moderated, and potentially suffering long term effects because of it, we observe indications of hypervigilant behavior. As one user posted, purposeful modification of one's behavior to keep up with the algorithm can lead to feelings of exhaustion and defeat. In their comment below, they also highlight the effort, as well as sophistication of strategy and tactics, needed to keep up and promote their business, linking it to a game of chess.

> "I have no idea if I have been shadow-banned because I use the same hashtags over and over (which I am now afraid of using ones like "fairy-wings" or "potionbottle" because I have used it too often even though that is literally what my product is and should be hashtagged as such) or if my views have dropped from the algorithm changing AGAIN. social media shouldn't have to be like chess, I shouldn't have to work THIS HARD to get just the people who follow me to see my stuff. Out of 700 people how are only 20 of them veiwing my stuff? I really don't have the time in the day to obsess over this. Instagram is killing my small businesses and MANY others." (2020)

Users express burnout, exhaustion, pointlessness, and helplessness in trying to keep up with the algorithm.

> "Its not worth it at all, I regret spending so much time over the years growing my art account on instagram and spending money boosting posts. The algorithm has destroyed my reach to the point that I feel all the work I put in was for nothing." (2019)

Users describe their adversarial relationship to algorithm. It is common for users to employ language of fighting, battling, or playing games, such as *"All accounts are in a constant battle with the algorithm." (2018).* They also attribute omniscient or all-powerful qualities to the algorithm they are up against: *"It seems to help if you just play along with the Instagram overlords' control system" (2019).* Not only do users feel they are in a battle against the algorithm, but they describe the algorithm as something they cannot possibly win against, an authority which uses unclear rules to dole out punishments and rewards.

> "to me it sounds ridiculous. why not just tell the user if a post is inappropriate and FOR WHAT REASON and preferably before posting? so that the user can correct mistakes and create more valuable content? why not publish the rules? so that people talk about it like its voodoo and about "the algorithm" like its some internet god? who you have to please for her to like you and approve your posts?" (2020)

Another characteristic of hypervigilance is the unrelenting assessment of potential threats, and the inability to take a break. Users apply this to recommending post consistency, posting every day, and not taking a break from social media as it may hurt your account in the long run. However, users also believe that they are at risk for additional punishment if they post consistently after or during any kind of ban. As one user expressed, *"When I got out of shadowban, I was using about 20 hashtags in the first comment. A couple of days later I got shadowbanned again, although it only lasted for 1 day. I got scared so I posted without any hashtags for a while and now only include in caption." (2019)* More drastically, users may refrain from posting altogether for an extended period of time such as exhibited in this comment, *"Nope, I've have not posted in 2 months. Hopefully the shadow ban is uplifted. Its so annoying, the action block threshold is so low for me" (2020)*. However, this is in direct contrast with the other fear that taking a break will permanently reduce an account's visibility. Users feel caught between multiple, and sometimes conflicting, pressures to manage their account success.

While Instagram may not explicitly punish a user for lack of activity, beliefs surrounding the potential for algorithmic deprioritization may cause users to feel pressure to continue producing content in order to preserve their visibility. As one user expressed, *"I post twice a week and it seems to be the sweet spot for me. Taking a break is tempting, especially since the engagement drop, but I think youre right - you get forgotten pretty fast if youre not keeping up and feeding the algorithm." (2020)* Aside from losing engagement, others expressed that taking a break from posting may put the user at risk of algorithmic punishment, *"Id hazard a guess that returning after an extended break could trigger a shadow ban." (2019)*

In order to avoid a negative outcome for their account, such as deprioritization, lower engagement, or shadowbanning, we observed that users may establish a regimented posting routine or self-impose quotas. However, these quotas are unrealistic for many: *"...because I don't post everyday I also lose followers quickly, which always annoyed me. I solely post videos, so to pump out a video a day is nearly impossible." (2017)*

Overall, the narratives observed within the data suggest that there is a common fear that lack of activity on Instagram can lead to implicit punishments. This may incentivize some users to create content on a regular basis simply to avoid punishment. Hypervigilance can take many forms, and here we note the trends of hypervigilant behavior in response to what users *believe* about the algorithms. We explore this hypervigilance further by drawing parallels to insecure attachment [407], which are often characterized by a perpetual state of hypervigilance and nervous system activation [428]. We see signals of anxious, avoidant, and disorganized behavior as reactions to precarious and unclear algorithmic punishment.

2.3.11 Interpreting User Sensemaking Through the Lens of Attachment Theory

RQ2: How might users' concerns about, and explanations of, algorithmic precarity on Instagram be interpreted through the lens of Attachment Theory, and what are the implications?

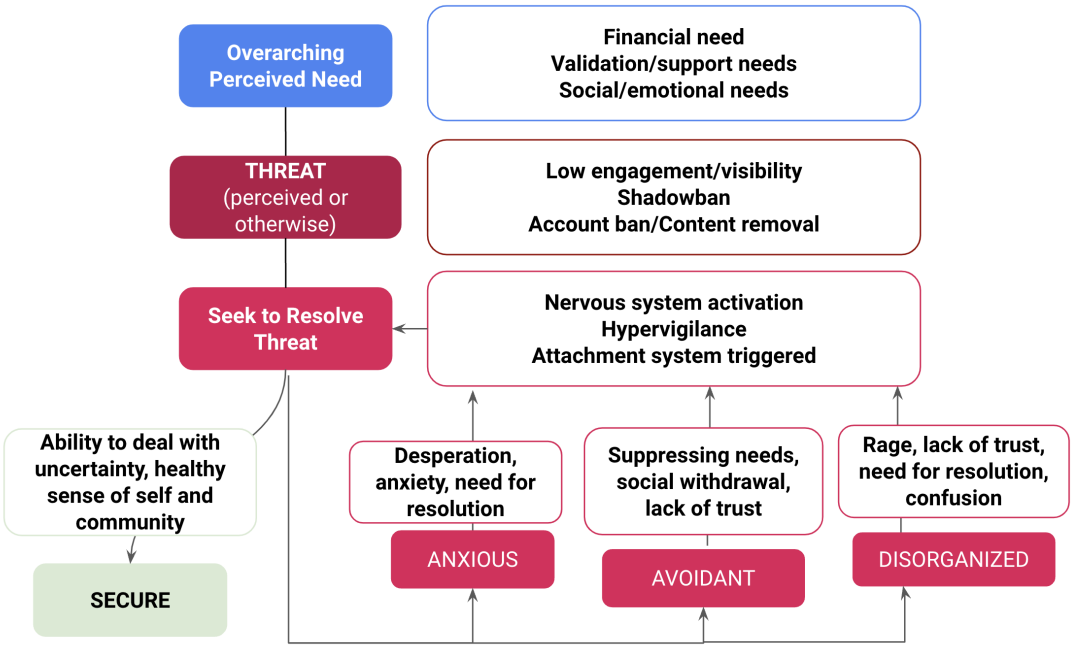


Figure 2.5: A proposed mechanism of insecure attachment responses in relation to social media algorithms. Adapted from Shaver and Mikulincer [428], a framework of the mechanism by which insecure attachments form. Key characteristics of the attachment styles are elaborated on in Table 7. Insecure attachment responses tend to stay in a perpetual activated state of nervous system activation and distress.

We explore user hypervigilance through the lens of Attachment Theory, with particular focus on how the uncertainty involved in algorithmic precarity contributes to an insecure dynamic between user and platform. When the authority on punishment and reward is inconsistent, the receiver of those consequences may respond in alignment with the three insecure attachment styles: avoidant, anxious, and disorganized. Users each have their own attachment predispositions, unique goals, and isolated experiences with algorithmic behavior, all affecting these outcomes. However, it would be unreasonable to diagnose, pathologize, or *prove* attachment styles from our sample; instead we use Attachment Theory [73] and platform paternalism [383] as lenses to explore how users experience algorithmic precarity, with a focus on community, financial stability, and self-presentation expectations. While a user may demonstrate an insecure attachment to Instagram with regards to their comments on Reddit, this does not say anything about that individual's relational attachment style. Instead, we use the lens of insecure attachment to describe the patterns of behavior that result from instability that users experience on Instagram, with the aims of both gaining insight into this dynamic as well as exploring the potential for fostering security. Table 7 summarizes these characterizations for how each of the attachment styles manifest in user behavior. Figure 2.5 adapts Shaver and Mikulincer [428]'s model of the mechanism of hyperactivation for the different attachment styles, but uses the algorithm as the source of threat. To answer RQ2, we demonstrate the various ways in which user's concerns align with the attachment styles, further discussing the implications in the Discussion.

2.3.11.1 Anxious Attachment

A characteristic of the anxious attachment style is the molding of the self to please an authority—sacrificing one's own needs in order to seek approval through a variety of strategies. Due to a lack of trust and stability, the anxious individual fears being abandoned or punished at any moment. In order to prevent these negative consequences, they preemptively try to adapt to

Attachment theory as applied to user reaction to algorithmic precarity on Instagram				
	Secure	*Anxious*	*Avoidant*	*Disorganized*
Key characteristics	ability to balance self and others, relies on community and also has ability to self-soothe, strong sense of self, authentic expression, emotional regulation [41, 379]	emotional activation, fear response, worry of losing connection, inauthentic expression to please others, high reliance on others for feedback [41, 40]	emotional suppression, disengagement, dissociation, distrust, lack of hope that needs will be met, loneliness, hyperindependence [41, 40]	combination of anxious and avoidant behaviors, aggression and rage, hatred mixed with desire for intimacy, dysregulation [370]
Manifestation on Instagram	seeks online community but not at the expense of their own authenticity or burnout, likely uses IG as a creative outlet, relies on other sources of connection or promotion, attempts to grow if in line with their personal boundaries, seeks compromise between user and platform	trying to follow all the unseen rules, attempting multiple courses of action, preemptive planning to avoid punishment, inauthentic expression to grow or be liked, worry and fear about the algorithm	leaving the platform, claiming it's "not worth it" to grow, stopping activity even if they express the desire to interact	anger at Instagram, aggressive language, continued attempts to reconnect and grow, desire to understand the rules despite frustration with them

Table 7: Characteristics of the different attachment styles and how they manifest with regards to the Instagram algorithm

whatever they think the authority wants. The anxious individual is *sometimes* rewarded or has their needs met, resulting in a continued cycle of repeating their coping behaviors [407, 428].

Anxiety often manifests as fear, as well as exhausting all possibilities to reduce uncertainty as quickly as possible [408]. In the following example, a user notices a glitch, describing their fear and *eight* different strategies they attempted in order to remedy the problem.

> "The same thing happened to me! More than 2/3 of my posts were gone this morning after the two factor authentication. Given all of our similar experiences, it's likely a glitch. I'm afraid to even post though, for fear that that would somehow cement my feed into its reduced state (as weird as that sounds)! I've tried so many different things to see if they work: I deactivated and reactivated two factor authentication, changed my password, logged out of all my IG accounts, deleted the app, restarted my phone, redownloaded the app and logged back in again, but it's still there. Like a few of you, I've also submitted a comment on the "report a problem" section and requested the data download. I'm curious to see how long this takes to fix!" (2018)

Another characteristic of anxious attachment is preemptive anticipation of another's needs. In order to maintain good standing, an anxious individual will try to prepare ahead of time in order to avoid punishment and possibly gain reward [407]. This user describes the Instagram algorithm as the "bane of their existence" but also discusses their goals of gaining favor with the algorithm. They are following the "rules" (which are unclear) but also trying to preemptively gain favor.

> "The Instagram algorithm is the bane of my own (and probably all of our) existence. So beyond trying to adhere to the rules and play nice Im interested in other ways that may make the algorithm favor my page more, which brings me to my question(s). Does watching the stories of people you follow and the sponsored stories between them help your standing in the eyes of the algorithm? Is there anything else I can be doing besides frequent content and adhering to the rules?" (2020)

Anxious attachment typically results in a lack of differentiation between self and others—the anxious individual becomes so focused on their partner or caregiver that they lose a healthy

sense of individuality. This is reflected in how users talks about "molding" themselves to each incarnation of the Instagram algorithm: *"The algorithm is an always changing brain so that means we have to constantly be molding ourself to the newer algorithm." (2020)*

> "This damn Instagram algorithm has been IMPOSSIBLE to keep up with, especially for my business pages. What're you doing to stay on top of it? It feels like we go through week-long waves of mastering it and get a ton of engagement, then we hit rock bottom again the next week and don't even get 100 likes or a single comment. HELP!!!!!!" (2021)

It is important to remember that anxious attachment is a valid and expected response to an unpredictable environment; it is a coping mechanism and in this case, a response to algorithmic precarity.

2.3.11.2 Avoidant Attachment

Avoidant attachment typically manifests as hyperindependence and disengagement from the source of unreliability. Key characteristics of the avoidant style are a dismissive and negative view of the authority, with the avoidant individual preferring to abandon the relationship rather than seek getting their needs met [335]. On Instagram, many people acknowledge the necessity of a social media presence for businesses and connection. For a user who feels compelled to stay on the platform, but also feels angry and insignificant, this incongruence may deepen feelings of dissatisfaction with social media. We observe that many users feel compelled to stay on Instagram even if they believe it cannot meet their needs. It is also important to acknowledge that users who exhibit avoidant behaviors may be less likely to post on r/Instagram; our observations here may underestimate the prominence of these behavioral manifestations. One user described how algorithmic changes are pushing people off Instagram, likening it to a 'slow death'.

> "There are a lot of assumptions here of foul play but I have A LOT of celebrity friends who basically abandoned Facebook last March due to their catastrophic algorithm changes. Now it appears the slow death is reaching Instagram. I even have a youtube watch partner who use to get 600k views a post that now completely left Facebook because now he sometimes get 100 likes on shit. Homie got 1.7 million followers... this is wild and dangerous." (2019)

When an avoidant individual's needs are continually not met, they tend to disengage and feel a sense of powerlessness. This is reflected in how users talk about their exasperation with Instagram and their intentions to leave:

> "Mine too, fuck Instagram I'm thinking about using other means - Facebook, YouTube. Instagrams algorithm and ghost banning is tiring me out " (2019)

> "This is something that really upset me about the feed sorting change. Before the change was implemented on my feed, which happened in mid-late march, I would scroll from the most recent post to the last post I remember viewing, which made it really easy to know when I was caught up. Since then I've almost stopped using Instagram altogether because I'm just not interested in viewing content the way they're telling me I should." (2016)

A key characteristic of avoidance is the perception that it is not worth the effort to get one's needs met—because the avoidant individual has observed enough evidence to discount that possibility [266]. Trust has been repeatedly broken and the only option for the avoidant individual is to disengage:

> "I ended up leaving Instagram for a while. Its been a couple months; that ban completely stripped me of all motivation." (2020)

> "It's as if the Instagram algorithm remembered everything I told it I *didn't* want to see and decided to show me just that :(I spent an entire weekend "hiding" (show me less of this) on every single one and it didn't "learn". It was never 100%, but it was nice to look at, now it's just memes, sports, makeup...whatever. At this point I've given up...I just completely ignore it. I don't think there's anything to do...they've done something to the algorithm :(" (2016)

Again, one must remember that displaying avoidant tendencies is not unreasonable. It suggests that the algorithm has consistently failed to show users how actions could change outcomes.

2.3.11.3 Disorganized

Disorganized attachment, unlike anxious and avoidant attachment, emphasizes a fear of authority figures in general [370]. In disorganized attachment, adults will employ a "confused approach and avoidance of the attachment figure for support and solace in times of need", as well as a higher degree of anger and hostility than is seen in anxious or avoidant attachment [370]:

> "Fuck instagram. They just deleted me without any warning. I censored all my pictures so they didnt show any nipples or anything. They don't respond to my appeals. It's not worth building a following on there when they just ban you so easily when youre making huge efforts to follow the rules." (2019)

> "Why does Instagram's algorithm fuck over talented creators while it somehow promotes people who lack talent? Why is it that every time I view the explore page, there are these annoying sexualized reels. The algorithm is going downhill for my art page no matter how hard I try. It seems like the only way to get the algorithm to work is through paying for promotions but I dont want to do that." (2020)

People with disorganized attachment have a lack of coherence in their mental representation of themselves, others, and relationships, which leads to confusion and conflict. One user recounts how their behavior on Instagram leads to a confusing reward and punishment cycle for their account:

> "Do you know what the weird thing is? I barely touch my IG account because of the drop in engagement, likes and followers. However, when I abandon it, the number of likes and followers starts growing faster than ever. However, as soon as I post to it again, the growth in likes, engagement and followers just stops. Once again, this adds credence to my theory that the IG algorithm is designed to act like drugs. They give you a taste, then when you want more, they cut you off and start demanding payment (promote). You end up paying more, getting back less and having to pay more and more to get less and less." (2020)

These users demonstrate a strong desire to follow the rules; they also highlight the ups and downs of their Instagram use as intense and seemingly intentional. However, they lack a coherent approach to their interactions, and incorporate both anxious and avoidant strategies.

2.3.11.4 Secure Attachment

Our data primarily showed signals of insecure attachment behaviors, but a minority of the comments responded in ways that aligned with securely attached tendencies. Secure attachment features a balance between self-reliance and community support, as well as a less activated response to external fluctuation. Secure attachment is also characterized by attunement to one's own goals and motivations that bolster a secure sense of self [464].

One user describes how they leveraged algorithmic knowledge to resist unrealistic beauty standards. Their sense of control, agency, and self-actualization is evidenced by their ability to cut out unwanted content and pursue their hobbies.

> "I actively stopped looking at beauty or fashion related posts on Instagram and unfollowed everyone that uses heavy Photoshop or facetune. Then I started to only focused on posts that had something to do with my hobbies. After a while the Instagram algorithm will stop showing you beauty related posts. I'm now starting a new art project because of a post I saw on insta :)" (2019)

Another user shows how they prioritize their own authenticity and fun, and resists what they think the algorithm expects from them, resulting in a discovery of a positive new hobby.

> "The algorithm was the reason I started collages. I was so annoyed by it that I decided to have some fun, instead of doing what they want me to do (which, I guess, is posting selfies and boring stuff)." (2019)

Another element of secure attachment is responding to uncertainty and precarity in a healthy way that prioritizes wellbeing. Below a user acknowledges frustrations, while also offering advice for growth. They mention trying to resist the pressure to grow, even though it can be a 'great tool'. They advise others to prioritize authenticity and happiness over growth, while still acknowledging the benefits of building a following.

> "Instagram algorithm is really messed up and even though your art is great, not many people may be able to see it. You can check some tips for growing on Instagram from other artist, try to engage with your followers and other fellow artist, or, if you want to, try some draw this in your style challenge for more exposure. Just don't feel so much pressured about this, social media is a great tool for knowing future clients and building a community, but never forget about your art and do what makes you happy ;) <3" (2020)

It is important to recognize that securely attached does not equate to happy. Securely attached individuals can and do navigate conflict and express displeasure with their circumstances. For example, users express frustrations while searching for a 'middle ground': *"I know I sound bitter and disgruntled. I know Instagram is providing a free service. I know I'm the product. But there has to be a middle ground, where Instagram can promote ads without *completely* ruining the feed." (2016)*

Secure attachment is also characterized by the ability to respond to uncertainty and discomfort with patience. Instead of a hypervigilant response to uncertainty, securely attached individuals can more easily self-regulate and weigh their options. We observe how users describe the importance of patience when dealing with fluctuations in the Instagram algorithm: *"First of all you should always check those things with multiple accounts and wait some time before jumping to conclusions, because there are fluctuations on IG." (2018)* or

> "Instagram algorithm treats you well as long as you keep people stay longer in the app. As long as you organically grow, youll see you get more engagement. It can take even 6 months or more of growth until you gain a nice amount of followers per day. Just keep on going, hope you the best" (2020)

Finally, we see how users can demonstrates a secure balance between authenticity and responding to algorithmic pressure. For example, *"I think the problem is probably algorithm. I'd love to fix the problem myself, but for now I'll just keep being genuine and hopefully will get the same in return :)"* (2018). As further illustrated, another user acknowledges the desire to grow, while also warning against sacrificing authenticity. They situate the center self-reflection and personal development. The author of this posts suggest using the decrease in engagement as motivation for self-discovery.

> "Im not recommending you turn into a bandwagon artist or falsely adopt color palettes/painting styles you dont like but know is popular with mainstream consumers. What Im saying is that a stagnation in activity could also be a sign to say that it might be time for new experimentation and personal development of your art. Algorithms are one thing but as artists, we should also consider that perhaps our activity might reflect our own feelings with our work." (2019)

These examples each demonstrate components of secure attachment that could be further explored by continued research, design interventions, community support, and discourse about social media and personal wellbeing.

Discussion

Our investigation reveals a variety of distressed responses to user-perceived algorithmic precarity. As evident in the language of Reddit data, Instagram users overwhelmingly view platform engagement through a punishment and reward paradigm, seeking to 'be treated well' and 'not get in trouble'. They also describe a helplessness and lack of agency to control their success and desired outcome on the platform. In this study of (in)secure attachment to 'the Instagram algorithm', we highlight the dynamic of unreliability and coping responses that arises due to precarity, opacity, and user beliefs. Here in our discussion, we ideate on the possibility of fostering secure attachment and consider interventions for supporting a more transparent and reliable algorithmic experience. We focus on potential pathways for promising HCI research endeavors that can design systems and evaluate the possibilities of fostering secure attachment to social media.

2.3.12 Limitations

Our findings demonstrate the impacts that algorithmic precarity has on users, in a way that mimics and/or heightens insecure attachment. It is important to clarify that Attachment Theory has primarily been applied to human relationships. However, we know that one's attachment style and subsequent behavior is greatly affected by the reliability of an *environment* as well [265, 433, 421], and we view the algorithm as part of a precarious environment impacting its users. We caution against any literal interpretation of Instagram as a parental figure, and instead use the metaphor to characterize the impacts of algorithmic precarity on user hypervigilance. Users already personify the algorithm (e.g. *'algorithmic overlords'*), or place their blame on sole individuals such as Mark Zuckerberg. Instead, we encourage users to direct their demands for transparency towards the human elements of algorithmic systems. In other words, efforts to resist these algorithms should be focused on steering oversight boards, challenging machine learning biases, calling for more insight into how the algorithms make decisions, critiquing Community Guidelines, and championing an improved appeals process.

However, when the uncertainty and precarity *cannot* be changed, due to the nature of ever-evolving trends and randomness, fostering secure attachment is necessary. Given our sample,

we suggest these findings are most applicable to creators on Instagram who already pay close attention to 'the algorithm', analytics, and trends—though the potential for fostering secure attachment may also serve teens [484], content consumers (those who do not post regularly or at all), and smaller accounts trying to connect with their in-person friends. More research is needed in specific populations of interest to explore the general applicability of our work.

2.3.13 Recommendations for Promoting Secure Attachment

We know that secure attachment represents a healthy balance between self and others, and the ability to more effectively handle uncertainty or fluctuation in relationships [464]. Secure attachment overlaps with trust, wellbeing, nervous system regulation, and self-worth. Insecure attachment can be transformed into "earned secure attachment" [379, 418, 124] and we draw from established interventions, as well as characteristics of secure attachment, to suggest possibilities for supporting earned secure attachment to precarious social media algorithms. We draw on both our empirical results and established elements of Attachment Theory to consider potential paths towards fostering secure attachment. For each recommendation, we outline how researchers could investigate and evaluate the efficacy of the approach.

However, we stress that social media alone is not responsible for manifestations of insecure attachment; nor can it be used in place of mental and physical health support. Larger systems of inequity must be deconstructed to fully support the wellbeing of social media users [338, 98]. This work simply provides a path forward for those interested in scaffolding secure attachment to social media—creatively envisioning future systems that foster and promote empowered digital connection.

2.3.13.1 User Agency via Goal Setting and Reflection

We observed that users felt disconnected from their own reasons for using Instagram, simply trying to 'stay afloat' and not be 'punished' as opposed to guiding their own goals for growth, connection, advocacy, or expression. The helplessness we observed demonstrates the lack of user agency and feelings of control over their own experience. In order to support user agency, a sense of control, and a more empowered way of interacting with social media, researchers could experiment with pre- and post-social media use prompts, surveys or other interventions that allow the user to explore *why* they use Instagram and what they gain from connecting digitally. Given that secure attachment often involves self-regulation skills and personal reflection [386, 124], we ground these recommendations in our data where users share preferred self-soothing strategies. We observe users reminding others to reconnect with the 'why' of social media use as a critical component of their personal experiences. This reflective process may allow the user to identify the behaviors they *do* like, and to focus their attention on serving their own goals as opposed to the algorithm. Reflection may be useful for guiding actions, bringing them in alignment with goals when scrolling. Possible design interventions may be to introduce quick touch points (SMS messages or links to short surveys) that prompt users for the type of content they want to see and how it supports their goals for social media use. In other cases, for example, if a user's goals are to connect with others, research could explore the efficacy of priming the user with a template of behaviors that are known to be associated with more meaningful connections, as opposed to simply viewing. It may be worth studying how to help users plan their time and accounts around the content they prefer to see; for example if a user prefers to see nature content they may make a separate Instagram account simply for plants. Each of these interventions should be evaluated for whether or not they affect hypervigilance and other signals of (in)secure attachment.

2.3.13.2 Scheduling Support Frameworks and Tools

Another proponent of secure attachment is reliable routine in the face of unpredictability [124]. Given the fluctuating nature of algorithmic success, it may be beneficial for users to control their own posting schedules and boundaries, while also engaging in self-regulation strategies[386]. One can imagine frameworks for helping users identify boundaries, such as not responding to DMs in the morning, or only spending a certain amount of time creating content before they deem it good enough to post. Hiniker et al. [232]'s *MyTime*, for example, contributes a framework for planning out social media *non*-use while supporting user agency and personal goals. We posit that interventions like *MyTime*, when applied to social media in combination with scheduling tools and APIs that allow users to batch their content ahead of time, may prove beneficial. However, a theme that emerged in our study was fear of long-lasting consequences for taking time off or missing a day of posting. While batching content ahead of time may circumvent this, we are also interested in how users *cope* with these fears and how they can be alleviated. Research to evaluate how promoting other forms of secure attachment affects these negative beliefs and anxieties around lasting punishments, or if there are ways to avoid those punishments with less effort from the user, hold great promise.

2.3.13.3 Peer Support for Modeling Secure Attachment

Our results saw users engaging in collective sensemaking to grapple with algorithmic punishments. We also observe users modeling secure attachment for others, such as encouraging them to "be patient" or focus on personal goals as opposed to pleasing the algorithm. While not as prevalent as expressions of distress, these responses highlight opportunities for peer support. We know that secure attachment involves self-reliance in harmony with community support [418]. How might we catalyze and promote peer support and modeling of secure attachment strategies and behaviors? Interventions could be as simple as an automatic bot on r/Instagram that prompts users to think about strategies grounded in secure attachment when it identifies anxious language. Peer support could also be initiated as a focus group, design session, or in a specialized app for users interested in building a more secure relationship to social media use. There are also possibilities for peer mentorship and support systems. Further research by the HCI community is needed to identify how to best support these needs, and to evaluate the efficacy of the suggestions provided among peers.

2.3.13.4 Providing Additional Context on Trends

Another way to support secure attachment is lessen uncertainty within the environment [200]. Users expressed distress with the current level of transparency, such as when content is moderated with no explanation. Additional transparency could reduce uncertainty and promote user security. While we may be unable to affect in-app features on Instagram, researchers could develop more detailed analytics that demonstrate natural fluctuations in trending topics, Instagram use, or prioritized features of the app. Oftentimes, a post may perform poorly simply because other content is trending or a particular hashtag is being overused that day. Content moderation may also fluctuate given how many reports are coming in, how many moderators are working, and due to new changes occurring on the backend. While access to these details is Instagram's intellectual property, researchers might invest in ways to provide enough insight to the user to help them identify natural or expected variability in visibility. Explanations of how different the system may behave day to day could be distressing or comforting; further research is required. These efforts could parallel others to increase data science and machine learning education in the general public.

2.3.13.5 Beyond Design

Investigating collective sensemaking through expressions on Reddit, demonstrate the continued and persistent distress that many social media users experience. We encourage localized community support for creative laborers to discuss and contend with algorithmic fluctuations, as well as mental health support that engages with lack of access and stigma for non-white, non-Western, identities. For each community affected by algorithmic precarity, support will look different. We encourage further participatory research into supporting specific needs of creative laborers with regards to algorithmic precarity and managing their responses to uncertainty.

Conclusion

Instagram users contend with algorithmic precarity in order to manage their financial opportunities, protect their privacy, avoid content moderation, connect with their communities, and present themselves in alignment with their self-presentation goals. Users fear lasting effects of moderation or poorly performing content, and refuse to take breaks for fear of additional algorithmic punishment. We find that users' hypervigilance in response algorithmic precarity mirrors insecure attachment. We detail the ways in which users demonstrate insecure attachment to platform algorithms, characterizing each style's manifestations of social media use. We suggest possible interventions for fostering secure attachment in order to promote user safety, stability, agency, trust, and wellbeing.

Verbatim Text

Verbatim text ends here.

2.3.14 Discriminatory Content Moderation

Verbatim Text

This marker denotes the beginning of previously published work, reference below.

Register, Yim, Grasso, Izzi, et al. *Beyond Initial Removal: Lasting Impacts of Discriminatory Content Moderation to Marginalized Creators on Instagram* In Proceedings of the CSCW ACM SIGCHI Conference on Computer-Supported Cooperative Work And Social Computing, 2024

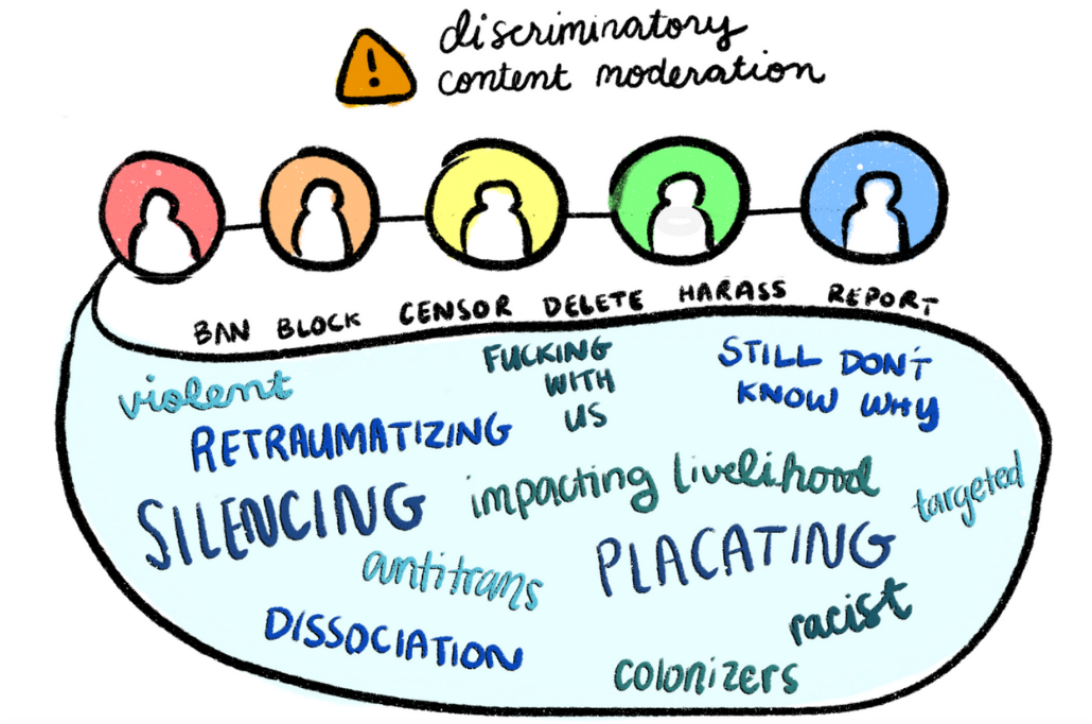


Figure 2.6: Original artwork representing the five narratives from marginalized content creators and some key words from their stories.

Beyond Initial Removal: Lasting Impacts of Discriminatory Content Moderation to Marginalized Creators on Instagram

Abstract

Recent work has demonstrated how content moderation practices on social media may unfairly affect marginalized individuals, for example by censoring women's bodies and misidentifying reclaimed terms as hate speech. This study documents and explores the direct experiences of marginalized creators who have been impacted by discriminatory content moderation on Instagram. Collaborating with our participants for over a year, we contribute five co-constructed narratives of discriminatory content moderation from advocates in trauma-informed care, LGBTQ+ sex education, anti-racism education, and beauty and body politics. In sharing these detailed personal accounts, not only do we shed light on their experiences with being blocked, banned, or deleted unfairly, but we delve deeper into the lasting impacts of these experiences to their livelihoods and mental health. Reflecting on their stories, we observe that content moderation on social media is deeply

entangled with the situated experiences of offline discrimination. As such, we document how each participant experiences moderation through the lens of their often intersectional identities. Using participatory research methods, we collectively strategize ways to learn from these individual accounts and resist discriminatory content moderation, as well as imagine possibilities for repair and accountability.

Introduction

Imagine you have spent years developing an Instagram page in support of survivors of sexual violence. You post recovery advice, mental health resources, and other timely content. The page is a community hub for peer support. You've built up over 20K followers, dedicating your time and energy to the page advocating against sexual violence. One day you log on to Instagram and there is nothing there. Your account is completely gone. No explanation. No information. You are completely in the dark.

This type of experience is common for advocates on Instagram (IG). They frequently face unclear and unreliable content moderation, often without access to a valid appeals process [340]. While we know that discriminatory content moderation is occurring, and prior work has documented specific cases [212], we know less about how the *experience* of being moderated impacts user's daily lives, access to community, and sense of self. Through participatory methods and documenting of experiential stories co-constructed alongside marginalized creators, we offer rich and deeply personal insight into the downstream effects of content moderation practices, with a focus on the lasting impacts to user wellbeing. In five case studies, we demonstrate the breadth of "content moderation gray areas" [212] as well as the situated experiences of marginalized creators; we demonstrate how moderation interacts with their identities, privileges, access, and mental health. Our findings illustrate the lasting impacts of discriminatory content moderation decisions, as told by creators themselves. Each story points to avenues for much-needed future research studying the unintended consequences of recommendations and polices of moderation online.

Scholars agree that Instagram has transformed into much more than a social networking site to connect with friends. In particular, many accounts are social advocacy and/or educational pages led by marginalized creators. For the purposes of this work, we view marginalization as a form of exclusion, oppression, and systemic lack of access due to perceived and constructed social difference in categories such as race, class, gender, and/or ability [189, 108, 107]. Systems of power and domination oppress across social difference in ways that are interlocking and multidimensional; and those who are categorized as different from the norm across any of these systems are marginalized, or systematically excluded – via policy, opportunity, or social enforcement. This work in particular explores experiences of advocates in trauma-informed care, LGBTQ+ sex education, anti-racism education, and beauty and body politics. Recent scholarship has demonstrated that marginalized users are often disproportionately affected by content moderation [145, 184, 527, 212, 24, 185, 182]; either directly via what is not allowed in Community Guidelines or by how those guidelines are deferentially enforced. For example, "female nipples" are not allowed to be visible according to IG's Community Guidelines. In other cases, LGBTQ+ social media users are mistakenly labeled as using hate speech when speaking about their own identities, e.g. using the term "bitch" as a term of endearment, referring to one's own trans identities, or using other reclaimed terms [141, 212]. Sometimes marginalized users re-post comments and messages they have received that contain hate speech, harassment, and bullying – these "callout" posts sometimes get taken down and yet the original comments do not, a stark failure of the system [340, 527]. Moderation can interact with one's marginalized identities; for example, Haimson et al. [212] conducted surveys to characterize and quantify disproportionate removals of content across Facebook, Twitter, and Instagram. They found that

political conservatives, transgender people, and Black people experienced content and account removals more often than others. For Black and/or Trans social media users, the content removed typically had to do with them expressing their marginalized identities.

Human-computer interaction (HCI), computational social science, design, and misinformation research (just to name a few) all contend with issues of content moderation on social media. In earnest attempts to avoid harm, research in these spaces often recommends uniform, punitive actions, such as restricting all nudity or quickly removing accounts found responsible for repeatedly spreading harmful content (as deemed by the platform) [263]. What we know little about, and is often ignored in these recommendations, are the downstream harms of these policies and recommendations. CSCW has successfully engaged in recommendations on providing more transparency [247], envisioning moderation beyond punishment and as opportunity for restorative justice [298], and the importance of discourse and education around community guidelines [249]. We contribute further insight into the effects of content moderation policies to vulnerable groups – furthering the scholarship in this area that seeks to engage with the difficulty of designing effective content moderation practices.

We define discriminatory content moderation as the disparate moderation of marginalized users by social media platforms. This could include, but is not limited to, increasing reach and engagement to users that hold normative identities in a way that is inaccessible to marginalized users, deleting or reducing reach and engagement of the content of marginalized users that are not violating Community Guidelines, and maintaining platform policies or guidelines that disadvantage marginalized users. This study contributes five situated narratives of content moderation, along with an important call to action for how we, as a research community, might structure priorities in order to mitigate harm to marginalized communities. Through participatory research and co-construction of narratives with our participants, this work aims to make space for these creators to accurately represent their *own* stories. Over the course of a year, we conducted interviews and follow-up narrative editing sessions with five creators, each with membership in and knowledge of specific marginalized communities, who have experienced discriminatory content moderation on IG. Each creator was recruited based on specific expertise which we elaborate on in Section 2.3.18. Their stories illustrate the enmeshment of online content moderation and offline discrimination and harm, with content moderation often reproducing larger systems of oppression in unseen and automated ways. In both our methods and discussion, we draw from Chen et al. [98]'s framework for understanding technology experiences through a trauma-informed lens, and explore various financial, emotional, and social consequences to being moderated – including exacerbation of trauma and violence. Further, we draw on Gerrard and Thornham [184]'s ideas of 'sexist assemblages', reflecting on how content moderation processes reproduce larger systems of institutional power[108]. In other words, we consider how content moderation can be seen as the online manifestation of other forms of policing that impact oppressed groups in complex ways.

We also build upon best practices for co-creation and collaborative knowledge production that has a long history at CSCW [70, 117]. Part of the contribution of this work is a demonstration of how to apply Chen et al. [98]'s principles for trauma-informed computing when evaluating the effects of a technology for research. We draw upon this work through co-written case studies that specifically demonstrate harm caused by discriminatory content moderation, conducting research using the trauma-informed guidelines for qualitative research that Chen et al. [98] build off of [512]. For example, one of Chen et al. [98]'s six principles for trauma-informed computing is Collaboration, described as "ensuring that trauma survivors are actively involved in decisions regarding their care and support. In essence, trauma survivors should be treated as 'experts in their own lives'[388], which means recognizing and valuing their opinions and decisions." For computing specifically, this includes "ensuring survivors have representation and input during the development and evaluation of new technologies". We employ the principle of collaboration,

along with participatory research and phenomenological methods, to arrive at the case studies presented in this work. Working with survivors of trauma, we also prioritized *consent-focused methods*, including content warnings, multiple opportunities for anonymity, full control over narrative presentation and the reporting of demographics information, and transparency through the process. We detail these methodological contributions in Section 2.3.16.

We deliberately chose to prioritize depth over breadth for this work, relying on relationship building with our participants to fully understand repercussions of discriminatory content moderation across multiple dimensions. Through particularizing these in-depth and specific narratives, we surface the details and *embodiment* of discriminatory content moderation on marginalized individuals. Through counter-storytelling [133, 132] and the situated knowledge [215] of our participants, we are able to gain insight into how discriminatory content moderation interacts with each of our participant's unique relationships to systems of power. Through particularization and participatory research we are able to identify gaps in current scholarship on discriminatory content moderation. We also explore the depth of experience through centering our participants and their needs in a collaborative process. We provide rich detail of the point of contact between a social media platform with billions of users and those who experience marginalization across multiple planes. In doing so, we illuminate how deeply entangled 'online' and 'offline' harms are, arguing that they are not separate entities but rather enmeshed together in embodied experiences of discrimination. This perspective was possible through close, trusted and long-term relationships with our participants, who divulged the traumatic impacts of discriminatory content moderation over the course of two years.

The motivating question for this work is as follows:

> **RQ:** How do the lived and situated experiences of marginalized creators provide insight into the lasting impacts of discriminatory content moderation on user wellbeing?

The five narratives documented here provide jumping off points for future exploration of discriminatory content moderation, specifically its long-term impacts and enmeshment with trauma and users' relationships to systems of oppression. While the roots of discriminatory content moderation stem from larger dynamics of oppression [182], we highlight specific affordances of the IG platform that compound harm in these cases. Finally, we, alongside our participants, draw the narratives and larger context of work in this domain together to suggest future pathways for accountability, repair, and transformative justice. Throughout, we present avenues for participatory and collective advocacy that centers marginalized communities, trusting their expertise of discriminatory experiences on social media.

Related Work

We are interested in the lived experiences and lasting impacts of discriminatory content moderation to creators with a variety of marginalized identities. We first review existing work that has illuminated the many ways that platforms employ automated and manual content moderation processes. This includes reviewing the history of content moderation practice, as well as considering the legal and social pressures for certain types of content to be moderated. Next, we cover scholarship that directly looks at how content moderation can perpetuate discrimination to marginalized groups. While prior work provides notable insight into both user experiences of moderation and the ways in which moderation can be discriminatory, we argue that a more in-depth look at these lasting impacts and how they are entangled with identity and power is beneficial for understanding this complex phenomena.

2.3.15 Discriminatory Content Moderation

Content moderation refers to "monitoring and vetting user-generated content for social media platforms of all types, in order to ensure that the content complies with legal and regulatory exigencies, site/community guidelines, user agreements, and falls within norms of taste and acceptability for that site and its cultural context" [405].Moderation actions include flagging hate speech, misinformation, nudity, cyberbullying, abuse, illegal activity, copyright infringement, and spam. Human moderators used to be solely responsible for content moderation on many online platforms, and often suffered psychological trauma from exposure to disturbing content [26, 51]. Today, many content moderation decisions are automated in some fashion; this switch to automation allows for increased scalability at a low cost [185]. Semi-automated content moderation is now the norm on most large platforms; in this approach, a variety of algorithms are used to identify undesirable content which is then passed to human moderators to judge. Here we focus on Instagram, which has one set of Community Guidelines applying to all content on the platform, as opposed to other platforms like reddit that allow for tailored intracommunity moderation, which poses different challenges [170, 427].

At Instagram, moderation decisions are made with respect to the Community Guidelines provided in their Help Center; these guidelines are the main source of information for users regarding IG policies. The Community Guidelines prohibit: nudity, purchasing likes/followers, copyright infringement, selling drugs, supporting or organizing hate groups, offering sexual services, selling regulated goods, blackmailing others, threatening others with 'credible, serious threats of harm', glorifying self-injury, promoting eating disorders, or sharing videos or images of graphic violence. Any user on IG can report another user's content or account via a built-in reporting option. If a user is on the receiving end of a report, they may receive a Violation in their account, usually specifying which Guideline was breached. The user then has access to a Violations tab to keep track of which content was ruled as violating Community Guidelines. With enough Violations, the user will receive a message saying *'Your account may be at risk for deletion'*. Typically, users can appeal Violations via the Help Center, but research suggests this is often unreliable or may not even appear as an option for some users [340, 212, 481]. Content moderation is not always a strict removal of content – some content moderation may consist of 'flagging' the content with a banner, pop-up, or click-to-view functionality. On Instagram, a photo may be blurred and labeled as 'Sensitive Content' and the user can click to reveal the image. Other examples of flagged content are redirection to the CDC for COVID-19 related content or Twitter's flag for potentially misleading or false information.

Recent work has highlighted the ways in which content moderation is discriminatory, with disproportionate removals of Black users, Trans users, activists, LGBTQ speech, sex workers, sexual educators, and the infamous cases of moderated female body hair or period blood [212, 61, 23, 24, 110, 305, 165, 141]. There is evidence to suggest "double standards" in content moderation, with certain similar cases treated differently than others [145]. This discrimination is sometimes attributed to some or all of the following elements: algorithmic oversights, human moderator bias, and the ways in which social media users take advantage of the reporting features on the platform. Errors such as mistagging or punishing in-group members for language patterns are often ascribed to improper training data or an inability of algorithms to interpret context [294]. Audits for machine learning bias can help us understand likelihoods of discrimination in automatic content moderation (e.g. see [394]), though it is difficult to generalize to the real-world effects and frequencies [56]. Furthermore, borderline content being treated as a violation of guidelines may also be the result of normative cultural values reproduced by platforms in ways that are nuanced and difficult to simply categorize as an "error" [301]. For example, the conflation of educational sexual imagery with prohibited sexual content could easily be the result of sex negative values as opposed to a "bug" or "error". We see that discriminatory content moderation can happen due to technical, legal, political, or social reasons, and is further

complicated by community reporting by users, with users engaging in 'report bombing' to target accounts and get them deleted [527]). It is also complicated by recent push to moderate misinformation, conspiracy theories, and rumors with particular regard to the information about the COVID-19 pandemic and COVID vaccine, as well as civic and electoral processes globally [31, 277, 412, 478, 328]. Research on discriminatory content moderation has focused on some of these various pressures, as well as the different types of content that gets moderated. Here we briefly summarize evidence of discriminatory content moderation on social media across various topics.

2.3.15.1 Moderating Hate Speech

Underlying many machine learning-based systems of automated content moderation are methods and tools for natural language processing (NLP), which can be used (with varying levels of success) to identify extremist speech, cyberbullying, or online harassment. For example, detecting hate speech and offensive language online in order to try to avoid abuse, racial slurs, and violence is an active area of work [494, 303]. Defining and bounding discriminatory content moderation is complicated by the complexities and challenges of attempting to moderate content at such a large scale. For example, hate speech moderation is used to remove text that *"encourages violence or attacks anyone based on race, ethnicity, national origin, sex, gender, gender identity, sexual orientation, religious affiliation, disabilities, or diseases"* (IG Community Guidlines). To detect and remove hate speech one must first agree on how hate speech should be defined [299] – should the definition include humor, factual statements, rely on additional context, or specifically target protected groups? These are active research questions.

Despite the difficulty of agreeing on a definition of hate speech, practical action is taken via automated detection and removal of hate speech on social media, prompting legal and social debate [510]. This automation gives rise to technical issues such as sampling and annotating the proper training data, attending to context, and managing the trade-off between false negatives and false positives. While high accuracy has been shown to be achievable on current hate speech databases, our work touches on the impact of false positives on marginalized individuals. To illustrate, consider recent efforts by many communities to reclaim slurs as a form of empowerment, e.g. the Black community reclaiming variations of the n-word or the LGBTQ+ community reclaiming "queer" or "dyke". Mozafari, Farahbakhsh, and Crespi [334] and Dias Oliva, Antonialli, and Gomes [141] found that automated hate speech detectors were highly biased against groups using reclaimed terms in a non-offensive way. Scheuerman, Branham, and Hamidi [422] also documents the various harms to trans individuals in technological spaces, noting how hate speech policies do not sufficiently engage with intersectionality.

Automated hate speech detection is limited, with several scholars questioning if it can ever be properly executed at scale [185, 341]. Olteanu, Talamadupula, and Varshney [369] describe how the mathematical costs may differ greatly from the perceived cost, harm, and impact to the user, and call for more human-centered risk assessment. Several scholars go further to imply that automated content moderation of hate speech and other content is actually exacerbating, rather than relieving, issues of content policy. This is due to the increased opacity of an already poorly understood process, as well as further complication of already existing inequities for marginalized users on social platforms [197].

2.3.15.2 Policing the Body

Social media platforms have a long history of policing images of the body, such as not allowing images of tampon strings or period blood, banning images of hair near the pubic area, and specifically banning "female nipples" [165]. Image classification is often used for the detection of nudity or 'sexual solicitation', though these are often conflated [165, 151, 197, 24]. In attempts

to remove sexual content, platform moderation has been shown to disproportionately targets sex workers, sex educators, queer models, body positive accounts, and even sexual assault survivor accounts [23, 143]. Prior work showed how LGBTQ+ content creators are often automatically categorized as "adult content" on Youtube, even if they are children's educators or making videos about something benign. If they mention queerness, it can be marked 18+, even if their content is not sexual in nature [455].

Scholars have also investigated social media's role in providing information on sexual health as social media is often the main source of sexual health information for minors. According to Borrás Pérez [65], the censorship of sexual health education and sexually-related content on social media platforms disproportionately impacts LGBTQ+ youth. Policies surrounding nudity and sexual content tend to punish those who do not conform to normative ideals regarding sex, sexuality, and nudity, while reifying bodies that do, hindering any non-normative expression of sexuality [184]. For example, Are [24]'s autoethnography work details how pole dancing is moderated on IG – threatening user rights, autonomy, and opportunity. Duguay, Burgess, and Suzor [151] investigates the platform values enforced via content moderation, community reporting, and content visibility, with a focus on queer women's experiences.

Black women in particular are subject to an overly disparaging gaze – with both automated systems and human moderators seeing them as more inherently sexual, threatening, and aggressive than white women [305, 212]. This perception of aggression is reflected in discriminatory moderation of what is considered *violent* – with anti-racism educators experiencing content removals when speaking out about the violence they themselves have experienced. Not only are racialized people at risk for being reported or harassed, they also conduct a lot of unpaid labor for the platforms – flagging racist comments and posts simply to keep their own community safer [434]. While marginalized creators experience death threats, rape threats, and other forms of abuse that may not be removed, their content can be taken down for anti-racist education [105, 312].

2.3.15.3 Violence and Harm

The moderation of violence and harm is a difficult task, especially considering the wide range of what can be considered violence [423]. Daz and Hecht-Felella [145] demonstrate this complexity and the 'double standards' of platform governance with a case study on removals of potential ISIS involvement vs. white nationalists on Facebook, stating that "rules against terrorist and violent extremist content remain opaque, failing to provide clarity on which groups have been designated as terrorist organizations and granting the platforms immense discretion in enforcement". IG bans images and videos of graphic violence, but prior work has shown that videos of police brutality often go viral [223, 128] – a traumatizing experience for many viewers of that content, despite its potential for raising awareness.

There are also different forms of violence, such as self-harm or eating disordered behaviors. IG prohibits the glorification of self-harm, including eating disorders, stating in Community Guidelines: *"Encouraging or urging people to embrace self-injury is counter to this environment of support, and we'll remove it or disable accounts if it's reported to us"*. Moderation is also dependent on how one defines an eating disorder, and it has been demonstrated that content moderation may actually reassert that 'real' eating disorders only look a certain way [168]. We know that many people use social media to disclose and discuss depression and mental health concerns; they may be at risk of being reported or flagged, especially when their speech is more negative and less positively received by others [296].

2.3.15.4 Community Reporting

Social consensus and community reporting play a large role in what gets moderated, sometimes in discriminatory ways. IG urges in Community Guidelines: *"Help us keep the community strong: Each of us is an important part of the IG community. If you see something that you think may violate our guidelines, please help us by using our built-in reporting option."* However, this opens the door for reproducing the very systems of violence they aim to reduce on their platforms. For example, people may report images of fat creators as nudity or sexual content, whereas they wouldn't report a thinner model wearing the same kind of clothing [144]. Zeng and Kaye [527] detail how "report bombing" can be used to target creators with marginalized identities. Content being removed for seemingly no reason or an obviously inapplicable reason may have to do with how users organize and take advantage of the system to report accounts en masse, even if a violation has not occurred [426, 105]. These instances cannot be disentangled from systems of power and domination, as accounts owned by marginalized users are more likely to experience this type of targeted reporting than those who hold normative identities [30]. Paralleling themes presented subsequently, Zeng and Kaye [527] report how *"Interviewees were exasperated that reporting was effective at removing their own videos that did not violate guidelines but ineffective at removing content they reported for being harmful, problematic and legitimately in violation of community guidelines"*.

2.3.16 Beyond the Point of Moderation

Prior work has documented the various ways in which content moderation can disparately affect those with marginalized identities. However, we know less about the lasting impacts of moderation and how the experience of moderation interacts with marginalization in broader contexts. Myers West [340] investigates user experiences of account and content bans on social media, and pay careful attention to the consequences of content moderation beyond threats to freedom of speech. They state: *"Although many users did discuss their experience with content moderation as an inhibition of their capacity for self-expression, the accounts surfaced a wider spectrum of consequences, some of which were particularly detrimental to users who are already in a marginal position in society"*. In this work, the authors focus on the financial, affective, and functional consequences that occur from content moderation, both online and offline – though their work was not specifically on marginalized creators. Furthermore, they are limited by short user responses and survey data restricted only to the point of removals/moderation. Haimson et al. [212] demonstrates a convincing report on the frequencies of discriminatory content moderation for Black, transgender, and/or politically conservative users, and gives brief insight into the perceptions held by those being moderated. For example, they demonstrate how one Black participant describes the censorship on Facebook as "upholding of white supremacy and racism". One of the examples provided by Chen et al. [98] for further investigation is Content Moderation, where they state *"removal itself could be traumatizing. Removing content shared within these communities can also hamper peer support by decreasing the information and resources available, and decrease safety by severing vital connections with advocates and resources"*. This prediction will be evidenced in our results.

Building on these important prior studies, our work delves deeper into the entire narrative of how one experiences discriminatory content moderation – from the initial point of removal to how it interacts with these larger systems, perceptions, their communities, and one's own experiences of their trauma as affected by technology and beyond. Alongside our participants, we carefully map out how every point of moderation is also a reflection of larger systems of discrimination and power, and how moderation continues to affect their lives long after the initial removal.

Methods

This study is designed to foreground the personal experiences of discriminatory content moderation, including the larger impacts that moderation has on marginalized individuals and their ability to do their work. Through interviews with marginalized creators we surface experiences and descriptions of each incident and provide knowledge that is embedded in that individual's identities, language, physical body, culture, and other experiences – an embodied and situated knowledge [411, 215]. In dialogue with our participants, we aim to gain deep understanding of their experiences through their stories, intonation, and gestures to illuminate the conditions of the phenomenon of discriminatory content moderation. Our research design draws on phenomenology, using disclosure of events as they appear to those who experience them [392]. Phenomenological studies aim to understand the subjective, lived experience of each of the study participants, and how the disclosure of these experiences can illuminate the conditions that allow for a particular phenomenon. We approach our participants' experiences with interpretive phenomenological analysis (IPA): "more likely to focus on how the whole experience is meaningful in the context of one's life as it has been, is being and might be lived." [443]

Phenomenological inquiry alongside our participants allow us as researchers to understand more about the situated and complex ways a phenomena like content moderation manifests for a marginalized individual [203, 323, 411, 456]. Participants are seen as co-researchers, and included in the crafting of narratives and analysis – with focus on the phenomenon as experienced by the participants themselves. Our participants each have a variety of marginalized identities that are central to their work. Their experience of content moderation interacts with their experiences as marginalized individuals, and the traumas they have sustained [8].

We also draw from counter-storytelling methodology[133, 132, 36, 322], grounded in critical race theory [135, 363, 306]. Counter-storytelling is "used to magnify the stories, experiences, narratives, and truths of underprivileged communities" in opposition to dominant narratives, with dominant referring to "practices, norms, and ideas that have the most power and influence in social, institutional, and economic structures"[92]). In other words, counter-storytelling is "a method of telling the stories of those people whose experiences are not often told"[451], which includes raced, classed, and gendered individuals [319]. Counter-stories have been successfully employed in legal contexts to give insight into things like "disparate impact" [134]. One such example particularly relevant to CSCW is Ogbonnaya-Ogburu et al. [363]'s *Critical Race Theory for HCI*, which provides several narratives to highlight ongoing problems of race in HCI. Solorzano and Yosso [450] points to the multiple functions of counter-stories, including but not limited to: building community and solidarity among those at the margins of society, challenge dominant perceptions and transform established belief systems, and allow us to envision and build a richer reality that takes into account both the stories and our current systems.

In our work, each of the participants' experiences can be viewed through a lens of intersectionality. Intersectional analysis [101] resists looking at oppressive forces in isolation, and instead regards the interaction of multiple vectors of oppression and privilege to explore an individual's situated and subjective standpoint [108].

Our approach goes deep into how our participants experience their own content, moderation of that content, and the downstream effects of content moderation practices and policies. These counter-stories each provide a window into a richer reality of how these practices and policies perpetuate harm.

2.3.17 Participant Recruitment and Timeline

We recruited participants who led an anti-oppression advocacy or educational IG account and had publicly talked about an experience being banned or moderated on IG in the year prior.

This moderation could apply to any user-generated-content, such as a post, story, or comment. One challenge inherent in working with vulnerable groups is to establish trust in the researcher-participant relationship. This is difficult without the researcher disclosing their motivations and prior experience. Therefore, the first two authors employed a recruitment strategy that was targeted; recruiting participants through direct invitation and word-of-mouth. They are creators that one of the authors had a prior relationship with, trusted their expertise, and knew that they had experienced discriminatory content moderation of various kinds. The two first authors are white, nonbinary-trans, and autistic, which was shared to participants following the feminist practice of reflevity and positionality statements. This undoubtedly had an effect on how participants chose to engage with us. Our research design, materials and methods explicitly addressed how we would be aware of, and counteract when possible, our own institutional privilege and the history of exploitative research on vulnerable populations [291].

We intentionally prioritize depth over breadth for this work. First, building trust and rapport takes intentional effort and time. Working with participants longitudinally allowed us to gain insight into the lasting effects of the content moderation experience. We explored how content moderation interacts with trauma [98]. Each of the narratives presented crafts a holistic view of an individual experience – something not captured by survey responses or frequencies of content removals. These narratives are not meant as a generalized view of discriminatory content moderation, though they do reflect some prior work's findings [212, 340]. Instead, they serve as a holistic and detailed view of this phenomena, as told by users themselves. Each story contains details worthy of further study, potential interventions for repair, and future opportunities for community-based participatory research.

Our aim was to recruit and interview at least five participants. The degree – both frequency and duration – of engagement with each participant limits the scalability of this approach, making larger numbers of participants prohibitive. We have worked with these individuals for nearly two years (*21 months*), continually facilitating updates to their narratives, our conclusions, and how each individual is represented in the text. We continue to work with several of the participants, including updates of how their perspectives have changed over time. The initial updates to the narrative writing spanned several months of back and forth editing, followed by contact every few months over the past year. At times, the first two authors portrayed the participants with incorrect wording or assumptions. We were able to repair these mistakes due to the interpersonal relationships the first two authors had developed with each participant, and offered more opportunity for edits so that participants could be portrayed in their own words. The entire process consisted of the interviews, several narrative iterations, a follow-up survey, reporting of our findings to participants, iteration on those findings with participants, updates throughout the process of submission and review, and edits to how their identities and demographics were portrayed in the text – all with reassurance that participants could edit the text up until the final camera-ready version. In this way, our approach differed from standard interview practice. We expand on the reasoning for this in Section 2.3.19.

We conducted semi-structured interviews via Zoom, with the two lead authors and one participant per interview. Initial interviews lasted 60-75 minutes. Participants received $ 200 as compensation during the process. We took care to offer financial compensation reflective of the rates that some of our participants charge for consultancy, and to compensate for the time commitment involved in this project.

2.3.18 Participants

Our participants were recruited based on their expertise in activism work as well as work providing education and community spaces in the pursuit of dismantling systems of oppression. They each run accounts with large communities, and engage with their community through support

groups, regular Live sessions, and other outreach. While providing participants demographics is common practice within research communities, we have found that this can feel reductive, often without consent of participants themselves. Demographic information is self-disclosed in the co-constructed narratives that follow. Here we provide an overview of the participants' work, as well as their Instagram accounts and their respective goals and audiences. It is important to note that we collaborated with participants for the following descriptions, with several iterations. In line with trauma-informed qualitative methodology [512], we collaborated diligently on how each participant is represented, translated, and shared in research texts.

Lauren is a certified victim advocate and trauma professional. Her account, MTMV Community Support Network, (mtmvcommunity) has 31K followers and centers the experiences of survivors of sexual violence through trauma-informed education and community peer support. Lilith is a model and activist, with her work centering the importance of representation and relationality. Her account, Lilith Fury | Goddess of Horror (lilith.fury) has 92.1K followers. She models for companies that do not have plus sized Indigenous and Latina representation. P3 holds a PhD in Black Studies and Women & Gender Studies. Her work focuses on the beauty and body politics of racism, and her account with upwards of 150K followers holds a mirror to white supremacy and whiteness. Constanza Eliana, an activist, writer, and educator, has 43.3K followers on the account Constanza Eliana | Decolonize (eliana.chinea). Through this account she works to decolonize wellness and self-care practices as well as providing anti-racism education, specifically centering the nuanced racialized experiences of being a Puerto Rican non-Black POC in the United States. Tuck, a holistic sex educator, shares their journey with gender transition, sex, and relationships with their 9K followers on the account Tuck Malloy (intra_sensual), along with queer and trans inclusive sex education workshops, tutorials, and information.

2.3.19 Trauma-Informed Methods

Working with survivors of trauma and those with marginalized identities, we followed trauma-informed protocols for qualitative research [512, 355], as well as the six principles for a trauma-informed approach [388, 355]. This included acknowledging our institutional privilege at the start of each interview, and reiterating our commitments to participant *collaboration and safety* [98, 388]. We designed our interview process with warm-up and debrief as suggested by Nonomura et al. [355], and reminded participants that they could skip any question, end the interview and still receive compensation, or ask us any questions about the research. Our consent model was detailed and customizable, with one participant asking to be involved to a lesser degree. Phenomenology and counter-storytelling methods allow the participants to represent themselves in their own words, and this is particularly important for trauma survivors who have experienced mis-characterization, silencing, or vilification. Participants asked to be in contact with others in our communications, and after obtaining consent from each person we shared contact details among the group, in line with the principle of *peer support*. Practicing the trauma-informed principles of *empowerment & choice*, we offered multiple and ongoing opportunities to edit one's characterization, as well as multiple opportunities for anonymization. We found that through the process of this research, participants realized new experiences (such as how much they had dissociated from their fears of being moderated or the compounding effects of harm), anxieties around how they were represented for research, shifting goals for their pages and communities, and even new violations on their accounts. The research process can take a long time, and with continual updates with the participants we were able to see that content moderation was still an ongoing issue throughout, as well as still a source of stress and concern. For *transparency*, we made clear our limitations that we did not have the power to influence Instagram's decisions or restore accounts/posts, but that shedding light on these stories

may be a step towards changes in policy and practice.

2.3.20 Interview Design and Execution

Interviews were divided into five sections of questions, each with a different goal and opportunity for relationship building with participants. **Introductions** gave the interviewers (the two lead authors) the opportunity to disclose their positionality. Interviewers reiterated informed consent before stating the research goals, emphasizing the participatory nature of the project. **Background** questions allowed the participants to share needed context about their online advocacy work. Participants were asked, for example, *"Tell us about your work on IG, including your audience, sponsorships, topics you cover, and the kinds of activities you engage in"*.

The next section, **Content Moderation Experiences**, delved into the participants' encounters with content moderation. We defined content moderation for participants as: the process of determining whether user-generated content adheres to the platforms community guidelines (policies) by both humans and automated systems and removing content that goes against these policies. After reviewing terms (content moderation, ban, algorithm, shadowban), we asked questions including: *"What was the most notable time you had something deleted?"* Notable may refer to the most memorable, most emotional, most financially impactful, most egregious, or most publicly visible. We collected information about which content was deleted, which actions they have had blocked, and how it affected them. We did not specifically ask about or probe for discriminatory content moderation experiences. The next section, **Theories About the Content Moderation Process**, investigated how our participants theorized about how content moderation works. However, in our interviews we learned that participants were not interested in the back-end mechanics of a harmful system, and did not want the responsibility of needing to know how it works. As interviews progressed, we eventually omitted most of this section, observing that attempting to involve our participants in the algorithmic underpinnings of the process came off as additional burden to them.

Finally, we concluded with a **Creative Cool-Down**. We acknowledged that the interview process can be charged and/or upsetting, and invited our participants to engage in some self-care questions and creative brainstorming. We asked about coping mechanisms for advocates facing content moderation. Next, we delved into the idea of repair and accountability from IG. We asked: "What actions, if any, could IG take to move towards repairing the harm they caused to you as well as to others more generally?". We posed a creative exercise: *"If you were to dream of your ultimate vision of an IG experience that is rewarding, validating, liberating, etc, what would that look like?"* It is important to us as researchers not only to uncover the experiences of harm, but to rely on participant narrative to collectively think about restorative design. We ended the interview by asking the participant if they could recommend anyone else for inclusion in the study.

2.3.21 Narrative Construction

The two lead authors separately listened to the audio recording of each interview, read the interview transcripts, and took reflexive notes – taking care not to analyze or generalize but to describe how the participant represented *themselves*. Interviews were automatically transcribed by Zoom software, but manually confirmed for correctness by the authors. In order to identify the main factors of each participants' story, the authors organized direct quotes and highlighted main themes and recurring descriptions of the experience. In writing the initial narratives, authors relied on both the interpreted main themes and direct quotes, in order to ensure that the narratives represented the data itself.

We engaged in a process of co-construction and verification for each narrative story by sharing

writing with individual participants, listening to feedback, and revising. We compared to the original transcripts to ensure the result reflected the interview dialogue. Authors stayed in contact with each of the participants for over 12 months, and sent editable drafts of their story to each participant. Participants were asked to complete a feedback survey which included questions about if and how the participants would like to be anonymized in the story and resulting research outputs, and how the written narrative needed to be revised. Participants responded, for example, with quotes they wanted amplified and interpretations that did not resonate with them. Participants edited their own stories, and the authors updated the narratives where needed after confirming the changes did not contradict the original interview material. Finally, authors distilled the main theme of the study to emerge collectively across all of the stories: Discriminatory content moderation is deeply enmeshed with broader experiences of discrimination, and that online and offline harms cannot be separated. Rather, they should be treated as informing and affecting each other, with discriminatory content moderation having far-reaching and lasting impacts, including but not limited to trauma, financial loss, and significant behavior changes.

Results

Below we share five personal narratives, co-constructed with each participant, and follow the inspired style of prior work Ogbonnaya-Ogburu et al. [363]. Each case study represents a counter-story [450, 133, 132], an account of the lived experience of being moderated online, and how that relates to race, gender, class, and other social categories of difference, alongside trauma. We rely on participants as experts of their own experiences [98, 388], and are concerned with their subjective realities and the impacts of those perspectives. For each of the named authors represented in these narratives, we encourage readers to seek their content directly for the most accurate insight into their work, either on Instagram or through their respective organization home pages.

The following narratives contain mention of sexual violence, racial violence, fatphobia, transphobia, whorephobia, PTSD, police killings, and death threats.

2.3.22 Story 1: Silencing Survivors

Lauren is a victim advocate and certified trauma educator, though she points out that survivors do their own advocacy – she is there as a facilitator, an educator, an activist, and a community organizer. She runs the IG account @mtmvcommunity, formerly known as @metoomanyvoices, describing it as "an account that speaks to dismantling rape myths, consent education, sexual assault, and trauma education. Simply put it's a community support network for survivors and supporters of survivors." In the summer of 2021, Lauren went to access the account, as she did regularly, only to find that it had been deleted. IG said that the account "violated Community Guidelines" and that the "account had been deactivated". She recalls this moment and describes it: "I literally had just been on there, like before my eyes, I got banned and kicked off and it just didn't make any sense what was happening, I thought for sure that it was a mistake, it was not a mistake". The experience was retraumatizing. "It was truly horrific. I just didn't understand what was happening. It was scary and I felt a loss of control ... trauma is you know, the absence of choice and control being taken away from you, I felt like my heart broke." Her community, her voice, her sense of safety, was all taken away in a moment. The community, her community, was rocked – their stability disrupted. Many of them rely on @mtmvcommunity; checking in on the account is the "first thing [they do] in the morning" for peer support. Not only did Lauren lose access to the community she had inspired and cultivated, but she feared she was being targeted – potentially by her abuser.

Lauren describes having to carefully navigate 'trolls' and violence on her page. As a credentialed trauma educator she is equipped to intervene when teens troll her page with rape jokes – but she wondered if her willingness to intervene by calling schools and offering sexual violence prevention education was what led to the deletion of her account. "It was extremely violating, whoever it was; knowing that it's an attack on me and my community, people that I care about and survivors who are already violated enough. Just talking about it makes me nauseous how violating." She points to the real impact of social media and content moderation: "It's not a game, and I feel like a lot of people didn't understand how serious and how personal and scary it felt. That's my livelihood, that's my voice being threatened and taken away." She emphasizes the personal toll as well, describing how she "barely slept the five and a half days that my account was down in total, I barely ate, I barely slept. I was a mess."

In response, Lauren quickly did a "deep dive" on the internet. She reached out to lawyers, press, and created a second account to bring attention to the issue. She asked members of her community to send reports to IG stating that @mtmvcommunity had been taken down unfairly, using the 'Something's Not Working' option in the Help Center. Her community members also supported her using other strategies such as looking up who to tag, finding information on strategies, and reaching out to people they knew that might be able to help. Just as suddenly as she lost the account, it was reactivated. She has still not been told which Guideline was violated or what led to the reversal decision: "I still don't know. They as an entity never acknowledged that my account came back, was taken down; there was no follow up".

Lauren also emphasizes the impact of her community support and how important that is for trauma survivors, "There are many mitigating factors when it comes to the impact and symptomatology of trauma, one of them being how supportive and affirming those around you are in the aftermath of traumatic experiences. It was a really traumatic experience that wrecked havoc on my nervous system at the time but the community came through for me in such an amazing way that I am not impacted by it anymore." She also highlights the juxtaposition between the impacts of this traumatic experience with with others, attributing her lack of long term effects to her community support. "I have thought about that is the big difference between my other past traumatic experiences that I am still working through is that I had people show up and care here."

She advocates for more direct access to communication with IG. "I understand that there's billions, literally billions, of users, but I think that there needs to be a way that people can actually speak to the company, the fact that there is not that's just a red flag." Expanding on this, she mentions how this lack of access to direct communication impacts her work, and for those who work with trauma survivors more broadly. "There's a lot of times where I'll have people that I have no idea who they are threatening to harm themselves. I shouldn't say a lot of time, but there have been instances where I'm truly afraid for other people's lives or I want to report them or try to get them more serious help, and IG takes zero responsibility for any of that." Lauren imagines some form of repair; she wants to "[get] answers" about what happened to her and why. "I don't even want an apology, because I don't think it would be sincere. I want acknowledgement that peoples' livelihoods, both financially and emotionally, are tied to these pages".

2.3.23 Story 2: Fat, Not Nude

Lilith runs an IG account, @lilith.fury, centered on plus-size representation, fat liberation, and demonstration of clothing and product options for disabled people. As a model and activist, she "tries to work with brands that either don't have plus size representation yet or don't have disabled people working with them or Latinas or Indigenous people". Coming from poverty, she doesn't try to influence people to spend money on something they can't afford, but rather wants

to help people to see their options. The account is her main source of modeling income, though she shares stories of her life as a model, an actress, a mother, an autistic person, someone with lipedema, a STEM student, an advocate, a horror fan, a lesbian, and an online friend as well.

Lilith has had many IG accounts. In the beginning, she had a more personal account with pictures of her friends, her dog, and her at the beach. She describes how the personal "account got removed also for nudity which was really weird because it was just me and my dog – like mostly pictures of my dog and then there's a couple pictures where you could see my head. I didn't even want to show my neck because of really bad body dysmorphia". In another case, she had another account where she "started trying to get used to [her] body." "I'd have a couple of full body pictures and that one actually was a whole bunch of people just being racist and whatnot that's why I lost that account, they just spam reported me I guess."

As she became more comfortable with herself, she ventured into modeling and acting opportunities, but describes the need to be "under the radar" to keep her account. As a fat woman who "dares to exist" she is continually reported by both automatic content moderation and other users. Her accounts have been repeatedly deleted. She does not have access to Branded Content (a critical feature for her livelihood). She is paid less than more novice creators for videos with more views. Her body is scrutinized more violently even when she abides by the Community Guidelines. If she appeals reported violations, "it gets worse because not only do they ... remove that post anyway and tell me that I'm wrong, but then they're like well, while we're checking your account, we found a few more things so anytime you appeal it's like they retaliate against you, for daring to stand up against them". People flag her posts, referring to her weight as "self-harm"; they report her body as "nudity" even when she is fully covered.

Lilith is extremely careful to abide by Community Guidelines for fear of losing her account and her livelihood. She chooses the clothing she models, the brands she agrees to work with, and the way she poses her body with policy guidelines in mind. Whether it's a pair of knee length shorts, a "hideous dress without cleavage", or a selfie in front of a nude statue at an art opening, her posts are frequently reported and deleted. She talks extensively about how she is not given equitable treatment under the Community Guidelines policy – stating that they "don't mean jack shit". She compares her experience to the many thin, conventionally attractive, white models who are never taken down for nudity, even when deliberately breaking the Guidelines. She is targeted, threatened, slandered, harassed, and abused online – all of these actions contributing to the deletion of her posts and risk to her livelihood. She says: "I don't want to lose my page. I'm terrified to lose my page. Every time IG goes down, I freak out, like oh my God, this is it, this is the day it's all gone. This is the day I lose my source of income, this is the day that I lose any and all opportunity, this is a day that everything is destroyed. And that is such a shitty feeling to constantly live in fear that your entire livelihood is not safe. It's not secure, and it can be taken away at any time. Just because somebody doesn't think that your body is worth being seen or that nothing you say is important because of how you look."

2.3.24 Story 3: Policed and Placated

P3 is a former therapist and cultural creative. She holds a PhD in Black Studies and Women & Gender Studies, and her work focuses on culture, identity, racial identity, as well as body and beauty politics (e.g. colorism, hair politics, skin bleaching). She is a "cultural critical source trying to dismantle white supremacy through teaching about white supremacy and our lived experiences". She talks about how white supremacy as a phrase can automatically put white people on the defense – "when they hear *white supremacy* they think Ku Klux Klan and they automatically say 'that's not me', but it *is* you. White supremacy does not just mean folks who rock the confederate flag or lynch people from trees. White supremacy is about how this is the lens through which you engage this entire world, and your entire existence." A key element of

her story is highlighting the institutional racism embedded into the algorithmic systems as well as the decisions made by human beings within the platform, "The insulting piece is that you also have the audacity to come up with diversity statements, you also have the audacity to say black lives matter you also hear you know you have the audacity to talk about share black stories. You want a gold star for being a basically decent human being and you're not even there."

The story she shared with us began in the summer of 2020, which P3 refers to as the "Black Summer", with its increased white interest in Black pain. Protests against racism, and specifically anti-black police violence, broke out across the US, accompanied by sudden white urgency around racism, a more covert racist violence. Social media hashtags popped up overnight, including "amplify Black voices" and "share Black stories", P3 recalls. During this time, she was invited to a campaign where Black activists took over white influencers' IG handles as a means of amplifying the voices of Black women. What was overlooked by campaign organizers, was the fact that exposure can result in violence against people who hold marginalized identities, especially if they are outspoken about systems of oppression. P3 speaks to this fact in regards to her antiracism work on IG, "for a lot of folks they think exposure is something that we should all want and be excited about I am not that person primarily because of the work that I do, and I want to be able to talk and engage the way that I want to."

Prior to Black Summer, P3 had used her IG page to engage with her Black community – either through information about her ongoing work and projects or through relatable memes and celebration of Black. Following her participation in the campaign, her audience became murky and unwanted. P3 describes the experience as follows: "Prior to that summer... I had a better sense of who my community was. The exposure that came with the campaign brought lots of new followers. Gaining another 30 to 40,000 followers was not something that I celebrated, primarily because most of them were not my target audience. My audience and all of the work that I do, for the most part, is Black people and/or folks who 'get it' whatever 'it' is." She describes the "critical and sometimes comedic eye with which I look through" and points out that the sudden interest in anti-racism that brought white people to her page also brought racist violence.

Following the campaign, she watched "in real time" as she was reported by a white woman for encouraging violence in a humorous post that was immediately taken down. She describes the post as "On surface what you saw was a Black man sitting in the front seat, looking at the camera being cute while two people are fighting in the backseat. It's fucking hilarious. Because most of us know that these two people are either brothers, cousins or friends" ... "But you, a white lady, come on and see two people fighting in the backseat of a car and asked me why am I 'condoning violence'." She exposes the reality of what moderating 'violence' on IG actually looks like, "Somebody walking around with an AK47 shooting somebody, is that violent? Absolutely yes. You allow the videos of a Black person being shot by the police, strangled, whatever, we get to watch those videos over and over and over again on your platform, but two boys fighting in a back seat is violence? Somebody in there knows which violence is okay according to their standards." In response to our description of 'unfair moderation' she stresses that this language is categorically incorrect: "I don't even know if it is a fairness involved – there's no *humanity.*"..." I don't have the word but fairness is too light."

In her telling of the experience, P3 draws attention to the entitlement that white people feel to police Black people, "I was very vocal on my page about [white supremacy and racism] as well. That makes white people uncomfortable. You want me to be anti racist in the ways you think is appropriate and that's not how it goes." Specifically, the report feature is available for everyone on IG but she highlights how white people have access to use it as a weapon similarly to how white people use the police to inflict racist violence on Black people, "No, this is not the white lady calling the cops on a black woman out in the world, but she called the cops on me [on IG]. In this day, it is the same shit and it's not just the calling of the cops. It's the fact

that you believe you have the *right* to police me. You believe that whatever you say and see is wrong, is wrong. And guess what, there are people who will look at your white body and believe you before they asked me anything." "It's about the power and the privilege that is afforded whiteness irrespective of who you are. You have a voice. [IG] makes that very clear."

P3 also offers a broader critique of the platform,"[IG has] the capacity to make every update known to man on this app and you're telling me you don't have the capacity to engage this algorithm? It's because you don't want to – somehow it's *working* for you – say that.". P3 highlights how these outcomes reflect the differential access to humanity for users, "For an algorithm, your very definition of a human being is white. Everyone else is some level of diversity". She would rather them call the Community Guidelines what they are – rules. She draws attention to the fundamentals of community that are missing, "Don't fake like we're community. We're not. Because if we were you would engage with the community differently. If these things are in place to somehow protect us, or to show us that you care, then why can't I engage with you?"

2.3.25 Story 4: Don't Say 'Decolonize'

Constanza Eliana's experience of content moderation sheds light on the weaknesses of both automatic content moderation and community policing. She explores the concepts of *fragility* and how much anti-racism educators are expected to censor themselves to be palatable to white audiences. Her initial goals around health and wellness were to increase representation for people of color in predominantly white wellness spaces – her work on the account @eliana.chinea quickly evolved into anti-racist and anti-colonial education. As a Puerto Rican she directly experiences colonization from the US Government, and trying to share her experiences within the predominantly white wellness industry was not welcomed: "you can't bring social justice or politics into wellness; you shouldn't talk about it; you're going to lose students." Constanza Eliana and others argue that part of self-care, wellness, and healing is resisting colonization in all its various forms. This form of self-care is not separate from wellness, but rather an integral part of promoting well-being. In the beginning, she was careful to use "proper language in order to get people not to just see me as being aggressive or argumentative but to like really see that my experiences are validated by other people and validated by anti racist theory". Now, her anti-racist and anti-colonization education page addresses "the implications of colonization, and continued colonization, on different racialized identities, but also on different ethnicities and nationalities". Her content "is intended to educate the public around the experiences of marginalized identities and the experience of colonization."

As Constanza Eliana's content expanded to topics of racism and colonization, she started to get posts removed. The first was about Black Lives Matter, and "since then I've had at least 15 posts either deleted permanently or deleted temporarily.". In describing the posts she has had taken down she says: "definitely everything has had some sort of racial component to it, calling out whiteness, calling out white supremacy.". Constanza Eliana has a carefully curated community of educators, and has done work to hold people who are causing harm to her and her community accountable. One prevailing issue she highlights is that white people continue to profit and benefit off of anti-racist work – often in the place of a more qualified non-white educator. When Constanza Eliana drew attention to how white anti-racism accounts were misunderstanding concepts through their white lens and yet receiving book deals, she received major pushback. She was banned from using IG's Live feature for 2 weeks.

In sharing her emotional experience of content moderation Constanza Eliana discusses how it feels to be continually tone policed, gaslit, and silenced in the offline world. "Definitely the first reaction is anger because I'm really not doing anything wrong – because typically when you get banned or deleted or something like that it's because the platform thinks you've done something

wrong, you violated something. And, for me, talking about social justice is not a violation, if anything, it should be talked about more. It also kind of messes with your value as well my sense of self worth because so much of what I do is tied up in not just the educational component, but my experience. So if something is being banned or taken down or shadowbanned as violating rules, meaning you've done something wrong. Then, that means my there's something wrong with my existence; my inherent experience. so definitely the mental health component of it is very tied to it as much as I do a lot of work around internalized oppression". These emotional responses dictate how much she is able to engage and continue her educational work. "It makes it too much of a burden and it's no longer you know fun for me to engage with my audience and engage with particularly other people of color and their experiences when I constantly have to worry about an algorithm you know either showing it or not, showing it banning it not banning it deleting it not deleting it like it's just too much of a burden."

Constanza Eliana has herself reported encountered instances of harm on IG, including a video of a white man in blackface and death threats she has received. These reports did not result in action by IG and she was told the content "did not go against Community Guidelines". She says "it's a very strange experience and so very because you know that social media isn't real and yet your real life has been threatened". She theorizes about why her content gets taken down and takes as an example posts that have included the term "white supremacy": "if I had to guess I would think that perhaps the platform, the content moderators, are within IG and Facebook and think that it's the use of those words that are causing the hatred or the violence or whatever, instead of actively taking a look at the people who really are white supremacists and over covert ways that are actually causing the violence". She goes on to point out that "so you can see, the power of white supremacy is coming into play: where they still decide what is racist and what is not; what is hate speech and what isn't." While she might be able to get around this censorship by purposefully misspelling 'white', "Personally, I refuse to do that because I think it dilutes the essence of the education and what I'm trying to put out there: that 'white' is what needs to be censored, instead of *the action of whiteness* that needs to be repaired.". Bringing her experience and expertise back to the Guidelines she says: "I think the way in which they are writing the guidelines gives too much room for the wrong people to be censored. So whoever is actually creating the rules, the guidelines, the policies isn't implementing anti racism into the the structure of the policies."

2.3.26 Story 5: Being Trans Isn't 'Bullying'

Tuck (@intra_sensual) is a certified holistic sex educator for mostly a queer audience. They use IG as a form of building community, promoting their business, and creatively exploring different kinds of educational content. Tuck began doing sex education work while at university – hosting small workshops on topics of consent and gender expression. They received pushback from the campus administration and decided to move their advocacy work to IG, which later became a way for them to support their own education-based business after losing their job due to the COVID pandemic. Starting out, "A lot of it was really personal. Because I was also exploring my gender identity and was exploring queerness a lot, and so it kind of came from that route of I'm just going to share what I'm learning about and what I'm teaching about and hopefully some other people will resonate with this". As their account grew, "Slowly the broader realm of the Internet started trickling into my life. It was at the point where I hit like 1 or 2000 followers that things started to get more messy on IG. It didn't feel as safe and comfy as being like 'Oh, these are just my friends and and you know my other sex educator communities that are following me'. There was more transphobia and more people pushing back against things that I was saying. I started to notice myself getting more anxious about the things that I was posting or the things that I was engaging with".

Tuck's content can at times push the boundaries of what is allowed by the Community Guidelines, and they have had many posts removed from the platform. "there's just been so many times [getting posts deleted], I feel like I have lost track at this point. The ones that hurt the most are like – there was one where I was posting something just about being trans and just being in gender exploration. And I really had no idea why somebody probably reported that. There was no nudity, it was pretty PG. I think I was writing about maybe some affirming sex that I had had, but that wasn't like definitely wasn't the main focus. And that got taken down and I just was like wow this sucks because this doesn't seem that edgy to me." The Community Guidelines that get cited in their experience seem incoherent: "It's usually like hate speech, or harassment and bullying. Those are the two ones that I often get for why this is being taken down. And that makes no sense. Whenever I see that I'm like well, but this is just such an incoherent definition of what hate speech is – clearly IG does not have an awareness of actual harassment and bullying."

In response, and as a way to manage the emotional impact, Tuck described dissociating from the constant battle with IG's content moderation. They feel "desensitized" – "a huge aspect of me staying on IG and maintaining my mental health is very much blocking some of these things out". They point out how "the systems of oppression and power stay in power because we're just trying to survive. So we're dissociating from these things and just trying to keep our heads down, and I think it can be really important to bring those feelings up, because then you get to engage with your rage."

For Tuck, "so many things come down to trying to eliminate sex from so many platforms." They point out the importance of examining who is making these decisions, and the power they have to do so: "I don't even know what's going on at IG behind the scenes at all ... I would also be really curious to know more about what the demographics are of that team; how they're making these choices; who they're getting money from." Tuck's repeated experiences with content moderation inform ideas around accountability and repair, and include a transformative version of IG that is more consent-based, drawing on ideas of consent and care from sex. IG has an opportunity to "shift their whole platform setup to be more consent based rather than more censorship based ... I definitely could see it being valuable to have more kind of like consent based check ins with people as they're engaging with content. If you wanted to follow a page that had sexual content, having to maybe read a disclaimer – like this is going to engage with sexual content, you should be aware of it before you follow this person, if that is okay with you proceed." In imagining consent-centered design Tuck draws on and builds from their background in sex education and sexual safety, advocating for how repair might be achieved by embracing more consent-based values as opposed to censorship and punishment.

Discussion

The above narratives cover a variety of experiences with content moderation each giving insight into the lived realities of this particular form of institutional harm on the platform. These narratives uncover the complex and varied details of being moderated for ones body, race, trauma, or advocacy. We identify several dimensions of harm across these stories – financial impact, isolation from community, exacerbated trauma, privacy concerns, unpaid labor, and impacts to self-worth and self presentation. Participants tell us about losing access to sponsorships, unpaid labor of handling 'trolls', exacerbated trauma in the face of uncertainty and opacity, as well as the tradeoff between vulnerability and visibility. However, each individuals experience varies depending on their exact circumstance, relationships to power, and for what they were moderated. In presenting such different accounts, the variability of harm, the complexity and nuance of solutions and recommendations, and the larger systems of oppression at play

are evident. In reflecting of these narratives we aim to move beyond simply reporting on discriminatory content moderation to look at the phenomena as a whole, focusing our lens on how it interacts with larger concepts of identity, privilege, and oppression. Each narrative reveals how content moderation reproduces systems of power and domination, blurring the arguably constructed boundary between the digital world and the "real world".

2.3.27 Online and Offline Harm Cannot Be Separated

For Lauren (Story 1), we saw how the experience of losing her account specifically triggered traumas around being silenced, targeted, and lacking control. She described the inability to eat and sleep during her attempts to learn and try everything she could in order to get her account back. She also demonstrated persistence and self-reliance, describing herself as a "fighter", because that is how she has had to take care of herself in the past. The fact that she was given no explanation for her account removal mirrored her lack of agency and voice as a survivor. Being cut off from her online community reflects the isolation and vulnerability she experienced after sharing her story. Lauren's experience of content moderation was directly linked to her experiences as a survivor of sexual violence.

Lilith (Story 2) also demonstrated persistence and unrelenting self-advocacy in the face of repeated discriminatory content moderation. This reflects the necessity for her to self-advocate to get proper medical care for her disabilities. Accustomed to not being listened to, she has developed strategies to continually fight for herself despite abuse and discrimination. This pattern of survival mirrors how she continued to find ways to support herself despite numerous account bans. It is also clear that Lilith experienced lack of access to means of success (that others had). On IG, she did not have access to branded content or the same payment rate as other creators. As a disabled Indigenous woman she has faced offline lack of access to resources, and she points out lack of equal opportunity online. Finally, her body is scrutinized more heavily than thin, white, able bodies – both by algorithms and humans. This is true in the context of content moderation, given what is considered lewd or inappropriate, and also offline in the world by judgments and shaming of her appearance.

P3 (Story 3) points out how online reporting is another form of policing of Black communities. Black people are deemed more violent, threatening, aggressive, and dangerous – both online and off. While institutions may promote Black Lives Matter or Celebrating Black Voices, they still reproduce white supremacy through content moderation and other platform features. P3 would rather IG admit that their current algorithms and processes are *working* for them, than hide behind a placating facade. P3 had adapted a strategy of simply posting joyful memes and videos, including the video for which she was reported. This mirrors the expertise in the Black community and Black scholarship, which also has promoted Black joy as resistance and intracommunity solidarity in the face of oppressive online experiences [253, 339]. To have this joke used against her demonstrates that even Black joy can be seen as threatening. P3 chose to disengage from IG, demonstrating that in some cases institutional harm ruptures trust so severely that there can be no repair.

Constanza Eliana (Story 4) applies analysis of power to her experiences; as a non-Black woman of color, she is deeply in touch with the nuance of her identities. She describes how it is a "certain type of person who targets me", because most white supremacists tend to subscribe to a Black/white binary. She is able to view herself and her experiences through nuanced questions of privilege, identity, impacts, and politics. Her expertise in mental health and wellness also informs her strategies for coping with violence, while also illuminating how wellness spaces perpetuate racism. Constanza Eliana analyzes the technologies she engages with in a similar way to how she dissects racist politics. She identifies the avenues that allow for oppression and silencing, as well as the avenues that amplify more privileged voices. For example, she is able

to look at how banning specific words may be helpful to avoid threats, but also circumvents accountability. In describing algorithmic precarity, she alludes to the volatile nature of public political opinion. She describes how one day her educational posts can be banned, while another day might receive 10,000 likes if it is a trendy topic. This mirrors the damaging trendiness of anti-racism when it is convenient for white people. Her analysis of IG's systems of content moderation were rich with insight into US politics, white fragility, and the nuanced experiences of being a Latina anti-racism activist.

Tuck's (Story 5) experience illuminates the conflation of gender, sex, and perceived threat. As a queer & trans sex educator, their content can be perceived as aggressive, inappropriate, and dangerous. Tuck was more comfortable in accepting that some of their sexual content might be received poorly by more conservative viewers, despite understanding that sex is not threatening or inappropriate. It was being moderated, reported, or harassed for their gender expression that was particularly painful. Offline, trans people are often viewed as inappropriate, sexual, dangerous to children and society, and not allowed to exist in certain spaces [411]. The policing of trans people may even be presented as "safety concerns", despite a lack of evidence to support these claims [19, 411, 447]. Anti-trans legislation tends to rely on laws regarding this safety, such as with bathroom bills, similarly to how content moderation is painted as a neutral or positive protection. The underlying rhetoric of anti-trans violence is that non-normative gender expression can be categorized as sexual aggression, fraudulent, or dangerous. These assumptions often lead to high rates of violence against and even death, specifically of Black, trans women [411, 152, 447]. For Tuck, the most painful points of moderation were when their authentic self was perceived as dangerous, just as how trans people are often regarded offline.

Each of these stories demonstrates how the experience of discriminatory content moderation is not limited to the online event of removal, but is rather an embodied experience, interacting with one's own identities, vulnerabilities, privileges, and perspectives. The effects can be longlasting, internalized, and must be situated in the larger context of how marginalized communities experience oppression. Each instance of moderation also affects how creators choose to engage with their audiences and communities – often with fear of moderation affecting their authentic expression.

2.3.28 Content Moderation Related Affordances and Shortcomings of Instagram Platform

While it is crucial to recognize that harms perpetuated on social media platforms are embedded in larger systems of oppression, we identify some key features of the IG platform that exacerbate these harms to marginalized communities. Here we address our guiding research question through the *specific* technical affordances and shortcomings that our participants shed light on.

Community Reporting, and its unclear relationship to algorithmic moderation, surfaced in many interviews. While it is beneficial for users to have some form of reporting power on social media, Community Reporting (and other like tools) are vulnerable to coordinated action and manipulation which can be used to collectively bully, harass, or silence marginalized individuals. The forced choice categories presented to users making a report can also contribute to discrimination. For example, Lilith explained how people would report her as 'Self Harm' for being fat. P3 describes how 'Violence' has such a broad meaning, ranging from graphic violent imagery to someone taking offense to a humorous post. Lilith described an experience of being moderated for "sale of illegal or regulated goods" when modeling knee-length shorts for a company. Tuck is consistently moderated for "nudity or sexual activity" as a sex educator. Queer expression may also be automatically age restricted, or hidden behind a click-to-view "Sensitive Content" blur filter. Sensitive Content control may be an admirable endeavor, promoting user agency and control over their own IG experience, allowing them to hide, for example, guns/

firearms, nudity, or violence. All accounts are placed at a “Standard” level of sensitive content control by default, and the user may select “Less” or “More” sensitive content than “Standard”. However, what is considered sensitive content, and automatically hidden without creator or user knowledge, can greatly impact marginalized creators’ work. Constanza Eliana criticized this change when it happened, educating her community on the possibilities of her educational content being automatically hidden without her consent. The interaction between violations accrued and advertising revenue is also unclear, with one’s ability to use Sponsored content at risk. For creators who rely on sponsorships for income, losing access to this feature translates into loss of earnings. Violations accrued may also affect the likelihood of being moderated on each subsequent incident; Lilith recalls getting a Story taken down for posing next to a nude statue something she doesn’t think would have happened if she hadn’t already been moderated so often. Even this perception affects her willingness to show up authentically online, fearing lasting impacts to her financial position and opportunity.

Another main theme across participants is the lack of communication from the platform when a creator experiences content moderation. Lauren experienced drastic moderation in the form of account deletion, without any notification or explanation. Others seconded this sentiment, noting the IG Help Center as unreliable and sometimes antagonistic. While a global team of reviewers implies diverse perspectives, it is important to grapple with the fact that this is often exploitative, low-paid labor [186]. Moderators may also lack appropriate local context to accurately judge content, and incorporate their own biases into final decisions.

In order to keep their own communities safe, marginalized users engage in unpaid labor moderating their own pages. It is important to note that our participants often used the same features to do this, but had either opposing or incommensurable opinions about the utility of the features, highlighting that a "one size fits all" approach is not feasible. For example, our participants describe the complexity of being able to hide words in their comments, a feature intended for anti-bullying; Lilith hides threats of sexual violence, while Lauren may sometimes engage with threatening commenters as a form of public education. Constanza Eliana notes that some pages may hide words that hold them accountable, such as ’racist’ or ’colonizer’, restricting public accountability. Tuck relies on quick Blocking when they encounter harassment. Restricting words becomes especially difficult when the words are used differently between in-group and out-group. For example, many queer people reclaim terms like dyke, fag, or slut.

Content moderation of billions of diverse users is a difficult task. It is needed to protect from egregious harm, and to ensure legality and safety of users. However, particular affordances of the platform stand out in the stories above. We also recognize that there was disagreement among our participants who each have their own identities, experiences, and values. This work is a starting point and motivates continuing this line of inquiry without treating marginalized creators as a monolith, and also providing spaces and resources for collaborative advocacy and resistance [485].

2.3.29 Limitations

This research is limited in the work it can do with and for vulnerable groups from the positionality of academia [291]. As we understand that academia and research and publishing practices are all institutions and processes that reproduce systems of power and domination, they are limited in their ability to work towards liberatory goals. Furthermore, both lead authors are white PhD students at a predominantly white institution. We acknowledge the importance of member research for building necessary trust for robust knowledge generation through interview studies. In one group session, it was noted by a participant that they felt uncomfortable being with a group of white people as a person of color. This arguably influenced their participation in the

activities and therefore the data we obtained from them. While we co-authored with participants themselves, and shared other identities in common with participants, it is undeniable that future work should prioritize researchers of color (as funded researchers or paid consultants on the work). This study received limited funds and we chose to allocate them towards our participants.

Potential Pathways for Repair Accountability, and Transformation

2.3.30 Participant Recommendations for Repair

We were also interested in potential solutions coming directly from those experiencing the harm of content moderation. It may be the case that for some creators, the harm they experienced is beyond repair. This is because online policing is a larger reflection of systemic oppression far beyond social media, and trust has been repeatedly broken. Among our participants, the most common suggestion for accountability is for IG to interrogate their own practices, representation, and Community Guidelines. This is in line with Gerrard [183]'s six opportunities for feminist intervention for content moderation. Participants stressed that they wanted Instagram to acknowledge that content moderation impacts creators' livelihoods – both emotionally and financially.

Second, several of our participants requested transparency to users – explaining *how* their content gets flagged and *for what reason*, as well as *who* is making the policies or the final judgments. While we are seeing increasing pressure for platforms to increase transparency to researchers through data access, this same access is not applied to users. Suzor et al. [463] provides a framework for meaningful transparency, with recommendations for platforms to cite specific rules and reasons for violations, and Calleberg [87] stresses how users are much less likely to trust decisions from an AI. While scholarship continues to engage with these problems, there is still significant effort needed for progress to become apparent to vulnerable stakeholders.

Participants also mentioned the 'uselessness' of the Help Center, with no standardized turnaround or specific communication. Several participants wanted to "speak to a live human". Lilith urged for standardized payment for creators, after learning that another creator with less views and followers was receiving more money for a video. She also wanted the built-in ability to provide *evidence* for her case when she submitted appeals. Constanza Eliana critiqued the reporting categories, demonstrating the wide range of how 'Violence' or 'Hate Speech' could be perceived or weaponized. It is unclear if there are specific teams who respond to types of reports. Tuck explored the idea of a more consent-based model for viewing content [242] – where the user is informed the kinds of sexual content that may appear from following a creator, and specifically opts in, perhaps even deciding how much they'd like to see.

Several participants commented on the cultural shifts required for any meaningful change – indicating how Instagram's profit model must be working for them properly and that so much would need to change for them to truly be accountable. This would involve hiring more BIPOC, disabled, trans, and sex-positive people in positions of power, working alongside creators who have been harmed, as well as applying a trauma-informed lens to the impacts of their decisions. Finally, accountability and repair may need to take the form of financial compensation – for the emotional and financial damages sustained.

It is crucial to view repair not just as immediate retribution, but also as a means of establishing safety, resolving uncertainty, and as a way to emotionally validate the experiences of the person harmed [516].

2.3.31 Call to Action for Researchers

Research has the power to shape platform design, policies, and features. Researchers can also focus work and resources on supporting community resistance and overall wellbeing for creative laborers and advocates.First and foremost our work highlights how intertwined 'online' and 'offline' experiences of technology are today. We deconstruct the boundary between harms experienced online and offline reality – and instead insist that *how* someone experiences harm is enmeshed in their identity, history, and positionality. This means that studying discriminatory content moderation must also attend to lasting impacts to one's financial security, access to community, and sense of identity and authenticity. Therefore, any ideation around repair or policy must consider the holistic view of those being affected.

We encourage researchers to think carefully about content moderation recommendations and to consider potential downstream effects to vulnerable populations. For example, increased moderation of potential misinformation could also result in more account bans for activists, particularly members of marginalized groups. Considering the potential consequences of a recommendation, and the financial, emotional, social, and mental burdens that could result, is a critical component. Basing recommendations for technical solutions solely on knowledge about bad actors does and will continue to harm marginalized creators. We also highlight the need for work examining lasting effects of content moderation, both on the intended problematic issues and unintended consequences.

Future work should attend to the immediate harms and need for repair in marginalized communities led by the communities themselves. This involves trusting the expertise of the people being impacted, and centering their experiences. For example, interviews should be trauma-informed and well compensated. IRBs should allow for co-authorship, co-construction of narratives, and participant-defined terms of consent. We see great potential for research that directly serves those harmed by discriminatory content moderation – for example, publishing narratives (with consent and participation) in order to help others feel less alone. We also encourage collaboration with other fields of research that have contended with issues of power and violence for decades. This includes feminist theory, critical race theory, disability studies, and queer and trans studies.

Conclusion

This research investigates discriminatory content moderation beyond the initial point of content removal – and considers the larger impacts, messages, and consequences of moderating members of marginalized communities on social media and beyond. Through participatory and trauma-informed methods with Instagram creators, we present five co-written case studies of the lasting impacts of content moderation on those with multiple marginalized identities. We explore the re-traumatization that can occur from being moderated for one's identities and experiences. The narratives span topics of sexual violence survivor advocacy, beauty and body politics, anti-racist education, queer & trans representation, and sexual health education. Each of these counter-stories [133] demonstrates the long-lasting downstream effects of content moderation policy and practice. Our findings show how deeply entwined content moderation is with one's identities, communities, and personal history. We provide possibilities for accountability and repair, as well as outline opportunities for future research. We encourage future work that carefully considers the embodied experiences of those facing discriminatory content moderation on social media.

Verbatim Text

Verbatim text ends here.

2.4 Chapter Summary and Contributions

✓ **Summary**

- ✓ With the rise of AI technology, we have also seen egregious AI failures and cases of algorithmic harm. For example, Google Images labeling Black faces as 'gorillas'[207] or Amazon's hiring software that excluded female candidates [246].
- ✓ Some other noteworthy cases include:
  - facial recognition failures on Black and Asian faces [80, 1]
  - discriminatory recidivism risk assessment in the COMPAS algorithm [72, 495, 142]
  - medical risk algorithm determined Black patients were healthier than they actually are, underestimating how much care they actually needed in comparison to white patients [362]
- ✓ Many of the high profile cases of algorithmic harm seem to have relatively straightforward fixes – for example, train on more data, or ensure parity for outcomes between populations.
- ✓ However, there are many more cases of algorithmic harm that are less straightforward and perhaps more insidious and complex. For example, the ways in which content moderation or search engine results reinforce gender, beauty, and body norms or contribute to sexualization of women and minorities [305, 352]. Another example is how large language models ingest 'unfathomable' amounts of training data, potentially overrepresenting majority (hegemonic) views, sexist language, or even conspiracy theories [47].
- ✓ This chapter covers the above cases, while seeking to expand the view of what is considered algorithmic harm. By specifically pointing to re-traumatizing effects of algorithms, I widen the conversation on algorithmic harms and how and where they occur.
- ✓ One of my published works in this chapter details how the lack of transparency, explainability, and consistency of the Instagram algorithms exacerbates hypervigilance and attachment responses in social media users [400].
- ✓ Another of my published works in this chapter demonstrates five case studies on discriminatory content moderation – when automatic content moderation algorithms discriminate against marginalized identities and how those creators are impacted by such algorithmic decisions.
- ✓ This chapter covers high profile cases of algorithmic harm, as well as expands the definition by including more subtle, insidious, or compounding impacts from AI algorithms.

△ Data Science Tip

Thorough Data Science includes in-depth analysis of potential outcomes from an AI model. Considering the potential downstream effects of model outcomes is one way to engage deeply with the models we create. Register et al. [400] and [401] demonstrate the end-user impacts of AI algorithms on human lives, particularly on social media. We can use such narratives in the Data Science classroom to demonstrate potential errors, biases, or poor assumptions in the sociotechnical systems we create.

★ Contributions

- a non-exhaustive but extensive list of algorithmic harms across industries, technologies, and harm types
- a psychological framework of attachment trauma exacerbated by social media algorithm precarity and the impacts on user mental health
- five case studies of discriminatory content moderation, as well as a methodological approach for participatory research with content creators harmed by social media algorithms

Part 3
Beyond Ethical Guidelines: Integrating AI Ethics Effectively

Not everything that is faced can be changed, but nothing can be changed until it is faced.

– James Baldwin

Abstract

In order to effectively teach AI ethics, I propose that cases of algorithmic harm and AI justice must be embedded into technical instruction in ways that are: specific, prescriptivist, action-centered, relatable, empathetic, contextual, expansive, preventative, and integrated.

In an attempt to regulate AI technology and mitigate algorithmic harms, many AI Ethics Standards have been released from big tech corporations as well as government agencies. While all unique, these standards can be summarised through five pillars of AI ethics. AI technology ought to follow these core principles: *beneficence, non-maleficence, autonomy, justice*, and *explicability*[173]. While these principles serve as what I call 'ethical bumpers', many scholars point to their apparent *uselessness*[336] – claiming that AI ethics standards are not easily translatable into practice[331]. I summarise the perspectives of these scholars in terms of AI education, identifying ways that we can integrate AI ethics effectively into AI curricula. As it stands, "ethics" is often left to the end of an AI course, or designated to another course entirely [180]. I look at empirically grounded, successful ways of embedding algorithmic harm case studies into AI education, as well as data literacy efforts that revolve around social justice. I summarise a variety of position papers in order to contributerecommendations for teaching AI ethics concepts. I argue that cases of algorithmic harm and AI justice must be embedded into technical instruction in ways that are: specific, prescriptivist, action-centered, relatable, empathetic, contextual, expansive, preventative, and integrated. I contribute peer-reviewed research that demonstrates one successful approach to teaching the technical topic of recommender systems using situated learning and the user's own data to do so. I find that users can identify potential algorithmic harms while simultaneously engaging in mathematical concepts through using my web application [398].

The previous chapter exposed a myriad of algorithmic harms; spanning different technologies, industries, and types of harm. Many of these acute cases of harm have been remedied, such as scrapping the Amazon hiring tool, or an option on social media to hide alcohol related ads [345]. However, many responses to algorithmic harm have been less than satisfactory, for example Google's response to the gorilla incident was to make it impossible to search for gorillas rather than fixing the underlying algorithmic problem of image classification in Google Photos [198]. The exact protocols of algorithmic harm prevention, mitigation, accountability, and repair are unclear – though there are a variety of efforts today to approach potential standards of AI ethics. In the past decade we have seen increasing interest in AI fairness, accountability, transparency, and responsibility, as well as *AI Ethics Standards*. It is now common practice for tech companies to provide their perspective on AI ethics – such as Google's 'People +

AI Guidebook', Microsoft's 'Responsible AI', IBM's 'Point of View on AI Ethics', AWS' 'Responsible Use of AI and ML', Apple's 'Ethics and Compliance', and Facebook's 'Five Pillars of Responsible AI'. The European Commission also put out a document on Trustworthy AI in 2019, which many have referred to and built off of. Organizations like AINow exist to specifically examine the social implications of artificial intelligence.

In the midst of the generative AI boom, AI safety and regulation are more important than ever. However, many major technology companies have actually laid off their ethics and responsible AI teams in favor of resourcing generative AI projects [130]. This is not to say that there are not ethics experts on and ethical guidelines for GenAI teams, nor that responsible AI is not being practiced in other ways, but the public disbanding of dedicated ethics teams within these companies in this public away may send the message that such research and practice is not valued. This includes the very public firing of Dr. Timnit Gebru from Google in 2021, following the *Stochastic Parrots paper* [47] warning of the dangers of LLMs. Exclusion is also reflected in the lack of diversity on AI councils and safety teams, such as Meta's AI council consisting of only white men [146]. At the same time, regulatory bodies are answering the call to create policy around AI. In 2023, the UN Advisory Body put out a report on *Governing AI for Humanity* and in 2024, the world's first major act to regulate AI passed in the EU (*EU AI Act*).

A large body of literature suggests that AI Ethics Standards are still lacking in practice; either because they are too overarching, ill-defined, non-contextual, technically insufficient, not actually reaching practitioners, or not engaging with issues of AI Justice. Furthermore, literature supports claims that they do not prepare future practitioners to deal with less obvious or more subtle cases of algorithmic harm when most Data Science professionals will overwhelmingly be more likely to deal with these less obvious harms than life-or-death decisions, and therefore need to be trained to anticipate or spot subtle injustices and mistakes.

AI Ethics education suffers from some of the same problems as AI Ethics Standards, not preparing students with contextual enough or practical enough instruction to effectively prevent or mitigate harm. Not only is AI Ethics education often insufficient, research also suggests that AI Ethics content is either left to the end of technical curricula or separated into another course entirely, marking it as 'other' or optional [180]. The most promising avenue for successful AI ethics education is a result of *integrating* ethics into the AI curriculum [395], including concerns and examples of ethics throughout technical instruction rather than leaving it to the end in non-practical ways.

This chapter reviews current AI Ethics standards and AI Ethics education efforts in order to summarize both strengths and weaknesses of the approaches – it also calls attention to opportunities for improvement. My specific contribution to this space is to demonstrate strategies for *embedded ethics*, using a situated learning approach to promote student care and curiosity in contexts of algorithmic harm. By combining the learner's own experiences of algorithmic harm with technical concepts, we can teach both algorithms and ethics simultaneously, intertwined with one another while retaining student attention, empathy, and curiosity. I demonstrate one example of this in my published work *Developing Self-Advocacy Skills through Machine Learning Education: The Case of Ad Recommendation on Facebook*, where users learn about recommender systems using their own Facebook data and then identify potential risks and harms that might occur from such an algorithm. The results show that learners can increase comprehension of a technical system (User-Based Collaborative Filtering) while identifying potential risks of that system by situating themselves as the stakeholders of the algorithm itself [398].

The previous chapter presented a widened view of what constitutes algorithmic harm, including less obvious and more downstream effects of algorithmic decisions. This chapter presents the argument for *embedded ethics*, integrating these examples into technical instruction. The following chapters provide more empirical evidence for the embedded ethics approach (Chapter

4), and provide example curricula and strategies for doing so (Chapter 5).

3.1 AI Ethics Standards

Reducing algorithmic harms in any meaningful way will likely take a coordinated, multi-faceted, large scale approach. This approach will likely combine policy, law, self-governance from companies, grassroots resistance, improved education, and changing social norms. While ethical criticism of automation can be traced back to scholarly work in the 1960's [506], AI Ethics Standards released to the public are a relatively new phenomena. These days, large tech companies each have a statement on responsible or ethical AI, often with guidelines and featured projects that demonstrate their ethical AI efforts. Scholarly research has reviewed popular AI Ethics standards and recommendations, leading to a distillation of the core principles of ethical AI. For example, Floridi and Cowls [173] provide a "unified framework of five principles for AI in society", inspired heavily from bioethics but adapted for specifically for Artifical Intelligence. Likewise, Jobin, Ienca, and Vayena [251] contribute an impactful Nature article that distills the global landscape of AI ethics, landing on five similar principles that are common in AI ethics standards across the globe. They review documents from Google, OpenAI, Microsoft, AINow, OECD, IEEE, Intel, ACM, DeepMind, Accenture, and up to nearly one hundred other documents all with recommendations for ethical and/or responsible AI. From these documents, Floridi and Cowls [173] distills the core principles as the following: *beneficence, non-maleficence, autonomy, justice* and *explicability* [332]. The first four principles come directly from traditional bioethics [44], with *explicability* added to account for specific challenges that arise from AI systems.

Definition 3.1

Beneficence: promoting wellbeing, preserving dignity, and sustaining the planet, also expressed as "the development of AI should ultimately promote the well-being of all sentient creatures" or "prioritize human well-being as an outcome in all system designs" or "common good".

Beneficence represents the idea of using AI for good, either for the wellbeing of people or the planet. As you can imagine, the definition of "good" is subjective – protecting one's country with facial recognition technology could be seen as "good", whereas others would argue it is an infringement on privacy and has disparate impacts to people of color. It may be seen as "good" to moderate harmful content online, whereas others would argue it is an infringement on freedom of speech and has disparate impacts to women. It may also be seen as "good" to optimize for profitibility of businesses to avoid unemployment, or to promote productivity and efficiency with AI. What is "good" can be argued based on values, and I cannot provide a standard for what is "good" any more than I can say what is objectively ethical. However, a consensus of "AI for Good" seems to focus on sustainability, education, or healthcare applications. For example, Google's *AI for Social Good* features projects like detecting retinopathy, forecasting floods, wildfire information resources, promoting literacy, disability support, creating greener cities, and healthcare applications. Microsoft's *AI for Good* mentions wildfires, sustainability, rainforest preservation, and breast cancer detection. IBM's *Data and AI for Social Impact* stands out from the sustainability and healthcare focused "good" by featuring projects that partner with non-profits to detect criminal justice inequities and deliver hygiene products to low-income families. While there may be no objective view of what is "good", beneficence is important as a *guardrail* – demonstrating the value of ensuring that AI is used in helpful and beneficial ways, even if what is "beneficial" is up for interpretation using ethical theory.

Definition 3.2

Non-Maleficence: privacy, security and 'capability caution', the concept of "do no harm", also expressed as "avoiding misuse" or "working against the risks arising from technological innovations".

Non-maleficence is the concept of "do no harm". While beneficence focuses on applying AI for good, non-maleficence tries to ensure that AI does not create adverse outcomes. Just as "good" is difficult to define, so is "harm", as discussed in the previous chapter 5. However, the subjectivity is not an excuse to disregard the concept of non-maleficence for AI systems. Non-maleficence may refer to mitigating privacy risks, and ensuring appropriate safety checks and constraints on the technology we create. We see a rising interest in AI Safety as well as prompt engineering for ChatGPT-powered technologies that ensure 'safety' and 'appropriate guardrails', demonstrating non-maleficence in action. As discussed in the previous chapter, harm from AI systems can occur for a multitude of reasons, and it is unclear whether non-maleficence encompasses them all. Floridi et al. [175] mentions that maleficence can occur from both 'inadvertent overuse' and 'intentional misuse', and both sources of harm should be taken into account.

Definition 3.3

Autonomy: the power to decide, also expressed as how autonomous systems "must not impair freedom of human beings to set their own standards and norms" or that "humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives."

The concept of *Autonomy* revolves around humanity's right to decide when it comes to how AI technology is used. This can mean deciding how and when AI technology is deployed, but also when it ought to be reversed or revoked (Floridi and Cowls [173] gives the example of a pilot being able to turn off the autopilot and regain control of a plane). In the creation of AI systems to take over human decision-making or creation, we willingly sacrifice some autonomy. Therefore, the principle of autonomy refers to the ability to choose how one's autonomy is ceded, as well as the power to reverse that sacrifice and regain human decision-making at any point. One example of prioritizing autonomy is Burrell et al. [84]'s work *When Users Control the Algorithms*, where they explore the kinds of autonomy that users desire over their information feeds on Twitter. Another example is that of human-in-the-loop AI, where human decisions are an integral part of the training process, such as in annotation of data and immediate feedback from humans to improve model performance [515]. Autonomy may also refer to one's data ownership rights, and the ability to "opt out" of an AI system, like AI resume analysis or being stored in a facial recognition database. Autonomy can refer to any levels in the hierarchy of AI deployment: from autonomy for the end-user to humanity's right to refuse AI integration for certain tasks (*i.e. medical care, policing, education, etc.*).

Definition 3.4

Justice: promoting prosperity, preserving solidarity, and preventing unfairness. Also expressed as "the development of AI should promote justice and seek to eliminate all types of discrimination".

In bioethics, *Justice* typically refers to the equal distribution of resources, such as medical treatments or access to care [44]. For AI, it is a bit less clear. In part, it does refer to "fairness", with a large body of literature devoted to just that [316, 35, 158, 364, 432, 236, 444, 396].

Mehrabi et al. [316] describes fairness as: "absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics. Thus, an unfair algorithm is one whose decisions are skewed toward a particular group of people." This view of justice focuses on non-discriminatory algorithms, that do not disfavor a population on the basis of their identities alone. However, the AI ethics standards reviewed in [251] and [173] also refer to the principle of justice as ensuring that AI benefits be distributed equally, which may or may not have to do with discriminatory algorithms. Later in this chapter I will summarise what Munn [336] calls the 'uselessness of AI ethics', and instead they put forth a new paradigm of *AI Justice*: taking into account the social and political dynamics surrounding the creation of AI technologies and how to consider AI through a lens of intersectionality. For now, though, *Justice* from the principles of AI ethics refers primarily to non-discriminatory algorithms and equal access to benefit.

Definition 3.5

Explicability: capable of being explained; enabling the other principles through intelligibility and accountability, incorporating both the epistemological sense of 'intelligibility' (as an answer to the question 'how does it work') and in the ethical sense of 'accountability' (as an answer to the question 'who is responsible for the way it works?')

Explicability is the fifth principle added by [173] to the four other principles from bioethics. Explicability maps most closely to ideas of explainability, interpretability, and transparency: understanding black-box models and making sense of their outputs [445, 518, 237, 211, 25]. Explicability attempts to directly encompass both explainability and accountability in one; making the assumption that if something is explainable it can be held accountable. Others agree with this sentiment, such as Arrieta et al. [25] describing interpretability as a means to "ensure impartiality in decision-making, i.e. to detect, and consequently, correct from bias in the training dataset", as well as "acting as an insurance that only meaningful variables infer the output, i.e., guaranteeing that an underlying truthful causality exists in the model reasoning". Smith-Renner et al. [445]'s work *No Explainability Without Accountability* makes this connection clear; that explainability alone is not enough and may even confuse users if the explanations are not being used to actually improve the models. *Explicability* serves as the principle of AI Ethics that allow the other four to be possible. When we *understand* the technology, we can ensure it is beneficent and non-maleficent, and we can make decisions about what should be kept autonomous and whether or not a technology is just.

The above definitions encapsulate the sentiment across nearly one hundred different approaches to AI ethics, distilling their similarities into these five principles. While these principles serve as high-level guardrails, their level of abstraction may be too vague for practical action [336]. For example, "beneficial to society" can be interpreted in many different ways, often with competing interests and different ethical foundations underlying what is considered "good" for society in the first place. This does not mean that a high-level framework of ethics is not necessary; each of these principles can be deconstructed, defined, and accompanied by practical recommendations. The following section attempts to show how such high-level standards are not immediately practical, but suggests opportunities for how they might be transformed into more useful forms.

AI Ethics Lack 'Teeth'

Some scholars claim that current AI Ethics Standards lack 'teeth': They are not legally enforceable, they lack specific technical implementation, there is a large gap between the standards and how to employ them in actual practice on real cases, and they do not meaningfully grapple with issues of oppression and social justice [332, 333, 403, 210, 209]. Consider this quote from Munn [336] in *The uselessness of AI ethics*:

> "These are *meaningless principles* which are contested or incoherent, making them difficult to apply; they are *isolated principles* situated in an industry and education system which largely ignores ethics; and they are *toothless principles* which lack consequences and adhere to corporate agendas. For these reasons, I argue that AI ethical principles are useless, failing to mitigate the racial, social, and environmental damages of AI technologies in any meaningful sense. The result is a gap between high-minded principles and technological practice."

Certainly these are bold claims, and this dissertation does not take the view that AI Ethics guidelines are completely useless. However, there is value in exploring Munn [336]'s claims further as they provide valid critiques and alternative avenues towards *AI Justice*. In referring to AI ethics principles as "meaningless", Munn [336] points to Mittelstadt [327]'s *Principles alone cannot guarantee ethical AI* – the principles tend to be too abstract, high-level, and vague, without specific action-guiding or practical recommendations. "Isolated" principles refers to the idea that the principles do not grapple with the larger ethical context in which technology is actually made: tech industry and tech education already suffer from sexism, racism, and other prejudices that result in larger ethics issues than just for AI technology specifically [367]. So while they are too vague in one area, they are also too specific in another. Finally, the claim that AI ethics principles are "toothless" refers to their lack of enforcibility, and how easily they can be overwritten by financial incentives [210]. It may even be the case that companies make such public displays of their (unenforceable) AI ethics efforts in order to avoid harsher regulation and policy by governing bodies [348].

One example of the "toothlessness" of ethics guidelines is seen in the null result from McNamara, Smith, and Murphy-Hill [314], who empirically tested the impact of ACM Code of Ethics on software engineering practices in 11 ethical vignettes (though not all specific to AI). One group received the ACM Code of Ethics, and the control group did not, and there was no observable difference in the outputs between groups. The authors ask "*if not a code of ethics, what techniques can improve ethical decision making in software engineering?*" This empirical result leads us to believe that we need something more than AI ethics standards or guidelines – what Morley et al. [331] calls "a pragmatic operationalism of AI ethics", expanding upon her previous work translating AI ethics principles into practices [332]. [331] points out that AI ethics guidelines are often spoken in "vague universals" and one way to improve operationalization is to lower the abstraction – give concrete examples and context. They encourage prescriptivism over mere diagnostics; instead of just pointing out that there *is* bias, give the developer next steps of what to do when there is bias. They encourage critical evaluation of moral responsibility beyond technical fixes, as well as honing the *judgment* skills of developers either in school or professional trainings. [332] describes how important it is to embrace uncertainty and deal with ethics on a case-by-case basis rather than relying on top-down standard rules. Engineers must be prepared to handle uncertainty and weigh tradeoffs along with stakeholder engagement – a practice not explicitly taught in most Data Science curricula today. Hagendorff [210] specifically calls for more ethics of care and a sense of social responsibility, eschewing checkbox-like guidelines for a deeper sense of empathy from engineers and practitioners.

> *Checkbox guidelines must not be the only "instruments" of AI ethics. A transition is required from a more deontologically oriented, action-restricting ethic based on universal abidance of principles and rules, to a situation-sensitive ethical approach based on virtues and personality dispositions, knowledge expansions, responsible autonomy and freedom of action. Such an AI ethics does not seek to subsume as many cases as possible under individual principles in an overgeneralizing way, but behaves sensitively towards individual situations and specific technical assemblages. Further, AI ethics should not try to discipline moral actors to adhere to normative principles, but emancipate them from potential inabilities to act self-responsibly on the basis of comprehensive knowledge, as well as empathy in situations where morally relevant decisions have to be made"*

Even broader, some scholars call for what they refer to as *AI Justice*, which looks at the wider sociopolitical landscape in which these technologies exist. AI Justice "reframes much of the discussion around AI ethics by drawing attention to the fact that the moral properties of algorithms are not internal to the models themselves but rather a product of the social systems within which they are deployed." [177] In other words, any discussion of AI ethics must be presented in context of the systems of *power*.

Definition 3.6

Relational Ethics: a decision making ethical framework that situates ethics in the context of relationships and centering marginalized and vulnerable populations. Related to both afrofeminism and feminist ethics of care [321]; *in context of AI:* "a framework that necessitates we re-examine our underlying working assumptions, compels us to interrogate hierarchical power asymmetries, and stimulates us to consider the broader, contingent, and interconnected background that algorithmic systems emerge from (and are deployed to) in the process of protecting the welfare of the most vulnerable. [57]

Birhane [57] describes this wider approach with the framing of relational ethics:

> *"Rethinking ethics is about undoing previous and current injustices to society's most minoritized and empowering the underserved and systematically disadvantaged. This entails not devising ways to "debias" datasets or derive abstract "fairness" metrics but zooming out and looking at the bigger picture. Relational ethics encourages us to examine fundamental questions and unstated assumptions. This includes interrogating asymmetrical and hierarchical power dynamics, deeply ingrained social and structural inequalities, and assumptions regarding knowledge, justice, and technology itself."*

Implied here is that the current approaches for mitigating or repairing algorithmic harm focus too closely on debiasing datasets and not enough on the broader context that shape the algorithmic systems we create. The problem of algorithmic harm is not just a technical one: but a political, economic, social, environmental, and educational problem. This sentiment is supported by [332] and [445] who demonstrate that approaches to AI ethics focus too much on *explainability* and *technical fixes*, without consideration of the broader issues of oppression, advocacy, and justice [177].

Of course, focusing on justice will still involve guidelines, regulations, technical solutions, education, and research. A focus on justice will not necessarily *solve* the algorithmic harms. But it is the foundation upon which we can build operationalized practices of addressing AI ethics concerns. There can be great value in technical solutions as well as regulation of AI technology – but the primary concern of this dissertation is a *cultural shift* as described by Morley et al. [333].

> *"AI practitioners (and indeed technologists in general) need to be encouraged to develop an understanding of the ethical implications of the products that they design by combining ethics theories in* ***mandatory courses provided to all data science, computer science, engineering****, etc., trainees... AI ethics researchers, in collaboration with journalists and public engagement specialists, should focus on making AI ethics* ***relatable*** *– both to AI practitioners and to the public. As we have discussed, highly abstract principles are potentially hindering rather than helping attempts to ensure AI products are developed pro-ethically. Bringing the principles down to a lower level of abstraction and focusing on more readily understandable questions such as 'is this the right solution for the problem?', 'is the solution working in the right way?', 'is the product having the right kind of impact?' can better support discussions about potential ethical implications."*

This brings us directly back to Data Science education, a pivotal period for future practitioners and a crucial point of intervention for a necessary cultural shift. While there are other points of intervention, such as professional development training and regulatory advancements, my work directly focuses on contributions to the education and preparation of future Data Scientists and technologists from high school through graduate studies. Intervening throughout a student's Data Science curriculum sets the foundation for more ethical considerations that are more *practiced, specific*, and *operationalized* than highly abstract ethical guidelines. Floridi et al. [175] also points to education as a key point of intervention for improving the efficacy of AI ethics in general. Their recommendation is as follows:

> *"Support the creation of educational curricula and public awareness activities around the societal, legal, and ethical impact of Artificial Intelligence. This may include: Curricula for schools, supporting the inclusion of computer science among the basic disciplines to be taught; Initiatives and qualification programmes in businesses dealing with AI technology, to educate employees on the societal, legal, and ethical impact of working alongside AI; A European-level recommendation to include ethics and human rights in the degrees of data and AI scientists and other scientific and engineering curricula dealing with computational and AI systems."*

This begs the question, how can we *effectively* teach AI ethics?

3.2 Recommendations for Effective AI Ethics Education

From what we know already, we can posit the key features of effective AI ethics education. According to Hicks and Irizarry [229], effective Data Science education in general ought to include *computing, connecting* and *creating*. Computing, of course, refers to preparing future Data Scientists with technical programming and statistical skills. As demonstrated in Chapter 1, many undergraduate curricula focus on such technical skills; mostly coming to consensus that students should be able to deal with large data, write code, understand AI algorithms, compute statistical tests, and communicate their findings [17]. However, scholars have argued that there is not *enough* computing built in to Data Science and Statistics curricula, which focus too much on theory and mathematics and not enough on programming [43, 229]. Hence, a core recommendation from Hicks and Irizarry [229] is to ensure there is enough education of *computing*. But it doesn't stop there; their next recommendation is *connecting*, a topic I have already explored thus far in this dissertation, and will elaborate further on in the upcoming chapters. *Connecting* involves using real-world cases and lowering the abstraction in what we teach. As discussed in my work [397], I argue that we often teach ML using either mathematical

Recommendation	Description	Example
Specific	lower the abstraction; rely less on mathematical notation and more on real-world cases [397, 229, 331, 333]	e.g. instead of discussing recommender systems generically, use a real world case study of dieting ads recommended on Instagram
Prescriptivist	provide recommendations for *action* as opposed to just diagnosing bias. "what are some options to do if X happens?"[331, 396, 394]	e.g. if a hiring tool is biased against women, more data needs to be collected from the minority class, and output hiring recommendations should also be class balanced
Action-centered	focus on the morality of *actions*, not *people*; avoid making students feel like they are being accused of being an unethical person[209, 210]	e.g. reporting a high accuracy score and omitting a low recall score is a mistake anyone could make but should not be done
Relatable	appeal to students interests, identities, and experiences; rely on situated knowledge to engage students [395]	e.g. allow students to choose their own data from sources like the UCI Machine Learning Repository
Empathetic	avoid "in-group favoritism" by also exposing students to cases they may not relate to and practice empathy-building via discussion or directed assignments[209, 229, 395, 245, 208]	e.g. content moderation algorithms help to remove unwanted content on social media, but end up causing lasting impacts of harm when they make errors
Contextual	discuss the social, economic, and political pressures surrounding the creation of technology in order to prepare students to engage with a wide variety of stakeholders and influences and make moral judgments on a case-by-case basis [210, 357, 395, 46]	e.g. facial recognition technology failing to work on Black faces is due to systemic racism and larger systems of oppression, and the solution to make these systems more effective for Black faces may increase unfair surveillance and overpolicing
Expansive	do not limit ethics cases to only life-or-death scenarios or the most egregious cases; expand the definition of what counts as algorithmic harm and expose students to different ethical frameworks other than utilitarianism [191, 210, 398]	e.g. labor rights, copyright law, and environmental impacts are also topics of ethics and harm
Preventative	ask students to practice *ethical foresight*[174]; imagining potential harms of a system before they happen and developing mitigation strategies; [394, 396]	e.g. perform red-teaming exercises in class
Integrated	do not leave ethics to the end of a course or only delegate it to a separate course entirely; include an ethical component as part of every case study and algorithm [330, 395, 531, 46, 180, 175, 42]	e.g. include ethical consideration for each lecture, perhaps by using Model Cards

Table 8: Recommendations for what characterizes effective AI ethics education

notation or abstract "toy" examples (e.g. x_1, x_2, x_3), or unrelatable and already cleaned datasets; and that this results in worse learning outcomes. Hicks and Irizarry [229] confirm this by arguing *"the typical approach in the classroom is to mathematically demonstrate that a specific method is an optimal solution to something, and then illustrate the method with an unrealistically clean dataset that fits the assumptions of the method in an equally unrealistic way"*, which they go on to argue hurts the student when they encounter real-world problems at the start of their careers. They recommend infusing the Data Science curricula with real-world cases as well as subject matter expertise and political context – as Data Science does not happen in a vacuum. Finally, they propose this idea of *creating*: where students must craft their own questions, agendas, and outputs rather than simply answering questions given to them. An integral part of Data Science is curiosity and creating new questions to find the answers to; and our current curricula often fall short on preparing students to ask their own questions of their data.

While the ideas of *computing, connecting* and *creating* did not explicitly refer to AI ethics, I believe we can apply the same logic. AI ethics can be discussed in terms of computing, such as teaching about algorithmic audits, fairness metrics, and overall model evaluation. AI ethics can be presented using real world cases that *connect* to the learner, ideally relying on their situated knowledge, experiences, and identities [431, 329]. And finally, students can be encouraged to formulate their own questions, be critical of technology, and create solutions to combat current algorithmic harms. Using current literature on AI ethics and computing education, I contribute the following recommendations for effective AI ethics education. AI ethics education ought to be: *specific, prescriptivist, action-centered, relatable, empathetic, contextual, expansive, preventative* and *integrated*.

We know that ethics guidelines seem to have little impact on the technology that students design and implement [314], which some scholars posit is due to them being too "high-level" and "vague" with no real action-guiding behind them (e.g. *'if this happens, then do this'*) [327, 333, 331, 210, 209]. Therefore, these scholars recommend that any effective AI ethics education ought to lower the abstraction by being more *specific*, and to give more prescriptivist examples of what one should *do* in the case of bias or harm, rather than just diagnosing that there *is* bias [331]. For example, Raji et al. [396] created an end-to-end framework for internal algorithmic auditing, outlining various *specific* steps in different stages of a project to mitigate algorithmic harm and ensure accountability.

Additionally, AI ethics education ought to focus on ethical *actions*, not ethical *people*. In other words, the ethics at hand should be discussed in an action-centered way as opposed to an agent-centered way [209]. An action-centered approach implies that actions can be unethical even if the people are not, and avoids placing blame on a student who may then choose to see themselves as "ethical" and therefore incapable of doing unethical actions. Alternatively, they may not want to be seen as an "unethical person" and therefore resist any AI ethics education that implies they are doing anything wrong. It is better to look at *actions* as potentially harmful or unethical, but acknowledge that anyone can make mistakes or face financial pressures that result in such harms.

Building upon the idea that cases should be *specific*, students may also benefit if they are *relatable* [397, 329, 377, 286, 395], appealing to student's real world experiences and identities. Raji, Scheuerman, and Amironesei [395]'s work *"You cant sit with us": exclusionary pedagogy in AI ethics education* explores the current state of AI ethics education by reviewing over 80 courses and they provide some recommendations specifically for including populations affected most by algorithmic decisions as well as experiential experts and community organizers. They even suggest what coursework might look like:

> *"We could imagine an AI ethics class, for example, in which the narration or participation of experiential experts is heavily featured, and students can share their own reflections on lived experiences with unjust algorithms, thus connecting*

personal testimonies to broader studied accounts of the social and ethical implications of deployed or speculative systems. The account of practitioners could be included in the class to explain the logics informing how technical artifacts are designed, and the kind of affordances or limitations present. Such exercises build empathy towards the variety of perspectives present on these issues, and encourage an open-mindedness to learn from and seek out an alternative lens" [395]

This quote attests to an important point regarding relatability, that Hagendorff [209] refers to as "in-group favoritism": we seem to naturally care more about issues that impact us directly, likely just part of human nature. Therefore, we must be careful about only introducing cases that are most relatable to a student – it may also be our moral responsibility to introduce cases of algorithmic harm and support empathy-building for the student to consider alternative lenses. Another component of effective AI ethics education is that it is *empathetic*. Luckily, it may be the case that through empathy-building, we can actually foster more relatability to the subject that may have been previously deemed unrelatable. As such, who was at first the "out-group" may become the "in-group" after an empathy building exercise, of which there are many. Janezic and Arsenault [245] identifies one of the most important tools for empathy-building is emphasis on shared humanity between groups, and that any accentuation of differences should be accompanied by a focus on similarities. I discuss more empathy-building strategies in Chapter Six.

Another way that AI ethics education could be made more successful is to situate it *in context* of larger social, political, and economic pressures surrounding the creation of technology, as well as preparing students to make moral judgments on a *case-by-case basis*, sensitive to the details of a particular instance. A *contextual* AI ethics education means engaging with larger systems of power, economic influences, and even workplace expectations to effectively prepare future practitioners to make ethical decisions in their careers. A student may want to make an ethical choice, but has not yet considered economic pressures from their company or communication strategies for making such a choice. AI education could benefit from providing larger context to the students, so that they can effectively make moral judgments on the job. Hagendorff [210] says:

Checkbox guidelines must not be the only "instruments" of AI ethics. A transition is required from a more deontologically oriented, action-restricting ethic based on universal abidance of principles and rules, to a situation-sensitive ethical approach based on virtues and personality dispositions, knowledge expansions, responsible autonomy and freedom of action. Such an AI ethics does not seek to subsume as many cases as possible under individual principles in an overgeneralizing way, but behaves sensitively towards individual situations and specific technical assemblages. Further, AI ethics should not try to discipline moral actors to adhere to normative principles, but emancipate them from potential inabilities to act self-responsibly on the basis of comprehensive knowledge, as well as empathy in situations where morally relevant decisions have to be made."

Nourbakhsh [357] writes a call to university faculty in which they highlight the importance and moral responsibility of integrating ethics into current AI curricula. The following illustrates another call for contextual case studies and opportunity for student judgment, with AI experts ideally situated to deliver such material.

"AI ethics is not the science of ethics, but rather shorthand for the notion of applying ethical considerations to issues surfaced by AI technologies: surveillance, information ownership, privacy, emotional manipulation, agency, autonomous military

operations, and so forth. As for integrating such reflection into an AI class, every case I am aware of does so, not with sage on a stage lecturing by the faculty member regarding Kant, but with case studies and small-group discussions on complex issues, lifting the students' eyes up from the technology to considering its possible social ramifications. No teacher can set the stage for such discussions better than an AI expert, who can speak concretely about face recognition errors, and how such mistakes can be inequitably distributed across marginalized populations."

Providing students with these specific case studies also allows them to explore a variety of harms, pressures, opinions, and solutions leading to my next recommendation that AI ethics education ought to be *expansive*. *Expansive* AI ethics education means expanding what is considered algorithmic harm as well as exposing students to different ethical frameworks. As discussed in Chapter 2, algorithmic harms come in a variety of forms. While it may be the case that the most egregious harms are in need of immediate attention, many algorithmic harms come from subtle or downstream effects from algorithmic decisions [400, 401]. It may even be the case that these more subtle or insidious harms are the majority of what Data Scientists will actually face on the job – with the life-or-death scenarios being fewer and further between. Exposing students to a wide range of algorithmic harms may allow them to mitigate harms before they occur, to look more closely for evidence of potential harms, and to find projects and ideas that personally relate to them. I discuss this more in the following Chapter. Another form of *expansion* is to move beyond Western ethical frameworks [534, 245, 210, 191] and to consider options besides utilitarianism or deontological rules.

Definition 3.7

Utilitarianism: the doctrine that an action is right insofar as it promotes happiness, and that the greatest happiness of the greatest number should be the guiding principle of conduct. *e.g. the classic trolley problem where an agent kills one to save five*

Definition 3.8

Deontological ethics: a normative ethical theory that the morality of an action should be based on whether that action itself is right or wrong under a series of rules and principles, rather than based on the consequences of the action. *e.g. the ten commandments*

Instead, students can consider case-by-case judgments sensitive to context, or explore other ethical theories such as Ethics of Care (EoC) or feminist ethics [353, 15, 442, 100, 187]. Ideally, students may mix-and-match depending on the context, the constraints they are under, their personal values, and the intended outcomes – but the point is to expose students to a broader view of both harms and solutions.

Definition 3.9

Ethics of Care: a moral theory that implies moral significance in the fundamental elements of relationships and dependencies in human life. Normatively, care ethics seeks to maintain relationships by contextualizing and promoting the well-being of care-givers and care-receivers in a network of social relations. Most often defined as a practice or virtue rather than a theory as such, "care" involves maintaining the world of, and meeting the needs of, ourself and others. It builds on the motivation to care for those who are dependent and vulnerable.

Through working with specific case studies and learning action-centered strategies for harm mitigation, students can also learn AI ethics in a way that is *preventative*: planning for and

mitigating harms before they occur. Floridi and Strait [174] describes this as ethical foresight analysis.

Definition 3.10

Ethical Foresight Analysis: a variety of analytical strategies for anticipating or predicting the ethical issues that new technological artefacts, services, and applications may raise.

Raji and Buolamwini [394]'s framework for internal algorithmic audits also provides strategies to assess systems for ethical risk *before* harm occurs. Floridi and Strait [174] point to strategies like red-teaming, looking at how a technology impacts different demographics, or specifically looking at how a system will or will not scale.

Definition 3.11

Red-Teaming: A red team is a group that pretends to be an enemy, attempts a physical or digital intrusion against an organization at the direction of that organization, then reports back so that the organization can improve their defenses.

Students should be given strategies for *preventing* harm before it occurs, and this will include developing students' critical thinking skills rather than strictly relying on checklists. The earlier concept of *creating* from Hicks and Irizarry [229] applies here: preparing students to creatively think about potential for harm from an algorithm or system. You may even position them as the 'red team' against each algorithm presented in class, an *action-centered* approach that helps them practice being part of the solution.

Finally, it is crucial that AI ethics content is *integrated*. As it stands now, AI ethics is either left to the end of a course or delegated to a separate course entirely. Garrett, Beard, and Fiesler [180] demonstrate that the majority of AI/ML courses include ethics only "if time allows" at the end. This stands to treat AI ethics as an afterthought, an optional and standalone component. Bates et al. [42] discuss some barriers stopping the integration of critical Data Science and FATE topics (*fairness, accountability, transparency, and ethics*; including limited time capacity for faculty, highly politicized environments, the general 'image' of Data Science, unclear metrics for evaluating FATE comprehension, and an overall underpreparedness and motivation for faculty to integrate ethics into their technical material. For example, the ethics components may even be left to one faculty member to "deal with the social part" [42]. However, AI ethics can be integrated as an integral part of the Data Science lifecycle, as demonstrated by Raji et al. [396]'s auditing framework that spans the entire lifecycle of a project. More companies are beginning to rely on *Model Cards* [326], a kind of Nutrition Facts for algorithms, which explicitly require the reporting of ethical considerations and limitations of a model, alongside its technical details. More scholars are beginning to call for *integration* of AI ethics throughout the technical curricula, with some successful results already [169, 531, 93, 509, 46]. For example, Matuk et al. [309] summarizes twelve different interactive lessons on data literacy that embed and promote social justice. Projects included using data to visualize pollution in students' own neighborhoods, students acting as data journalists to report on mental health issues or gun ownership data, data exploration of financial disparities across the city and situating their school as existing in a less resourced area, and several other projects where students used *their own data* to make meaning and relate to social justice topics. Students from underrepresented backgrounds use their situated experiences and funds of knowledge to engage in data visualization, analysis, and reporting – all through a social impact lens that grapples with environmental sustainability, medical care, mental health, financial opportunity, and racial justice. It is indeed possible to embed ethics in teaching Data Science and AI, and there are many calls to do so.

The following work represents one of my own published contributions to an integrated AI

ethics approach, building upon my situated learning findings [397] and focusing on a social media algorithm and personal data: a technical lesson on recommender systems that embeds questions of algorithmic harm in a situated way. It is *specific* because it provides a specific case study: Facebook ad recommendations via the user-based collaborative filtering algorithm. It is *relatable* because it uses the learner's *own* Facebook data to teach the concepts. It is *contextual* because it deals with a real system and real data and a real Facebook user looking at their own personal data collected by the company. It is *expansive* because it encourages the learner to provide their own ideas on what constitutes algorithmic harm, however small, and *preventative* because it asks them to ideate strategies for harm mitigation (either algorithmically or through their own behavior). Finally, it is *integrated* as it seamlessly combines technical instruction of an algorithm with ethical consideration along the way. This particular case study was more focused on helping users identify potential algorithmic harms, but could be adapted for engineers to mitigate those harms and building empathy for users while simultaneously learning the innerworkings and risks of recommender systems in the classroom.

3.3 Facebook Ad Recommendation Case Study

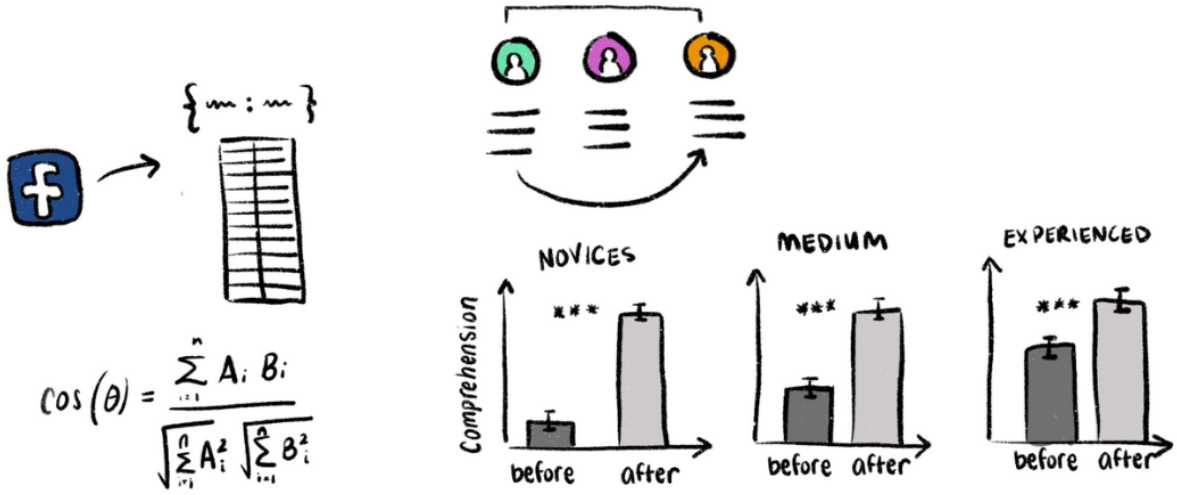


Figure 3.1: Original artwork depicting Facebook data, the cosine similarity metric, the custom tutorial described in the following paper, and a cartoon version of the results.

Author's Preface

I was inspired to build upon ideas of using personal data for machine learning education. In my previous work, I justified the ideas behind situated learning for ML, and how it affected both technical knowledge and self-advocacy arguments. Building upon this, I wanted to apply it to a real-world case study *in context*. There was a variety of cases to confirm that social media algorithms had the potential for harm, and personal social media data was something one *could* obtain given European Union General Data Protection Regulation (GDPR) regulation. I was particularly fascinated by the fact that despite having *access* to one's own data, it was unclear what was actually being *done* with it in practice. One of the core algorithms of social media is recommended ads and content – most likely powered by some form of collaborative filtering. The probe I built was meant to shed light on this algorithm (teaching some technical knowledge) as well as find out the relationship between that technical illumination and how a user spoke about algorithmic harm. We were also able to gain insight into the *types* of algorithmic harms occurring that we may not have thought of – confirming some of my ideas about dieting and alcohol ads that I had been thinking about for a while. Facebook has since dwindled greatly in popularity, but the core of what I built speaks to any collaborative filtering algorithm employed today. This study represents a step towards success for integrating situated learning, real-world context, technical knowledge, and critical judgments about algorithmic harm – all in one scalable, educational experience. While I certainly would change things about it if doing it again today, it represented an attempt at integrating all of these important pillars into an educational experience – that worked! To me, it represented the possibilities of enmeshing ethics in technical instruction. It also dealt with more contextual and situated versions of algorithmic harm than we may be used to hearing about in the news. For future designs, I would focus on probing for more generative ideas about what we could change algorithmically, ways that the design could work alongside users' best interest, how we should audit such systems, and more about the potential for harms that I haven't thought of at the time. I would also like to see users think *beyond* their own situated experiences, engaging with personas of another; weighing how their preferences might contradict with someone else's. I am also interested in opportunities for repair, and the kinds of avenues for accountability or self-protection after harm has occurred. This study also

fueled my interest in how the public engages with social media algorithms, the potential for social media algorithms to cause harm, and what users understand about these algorithms. This led into my work on both algorithmic precarity and discriminatory content moderation, each of which shed light on insidious consequences of algorithmic design (discussed in the previous chapter). I include this paper in the chapter on integrating ethics into AI education because this application served as one example of embedded AI ethics in practice.

Verbatim Text

Yim Register and Emma S Spiro. "Developing Self-Advocacy Skills through Machine Learning Education: The Case of Ad Recommendation on Facebook". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 16. 2022, pp. 817–828

Developing Self-Advocacy Skills through Machine Learning Education: The Case of Ad Recommendation on Facebook

Abstract

Facebook users interact with algorithms every day. These algorithms can perpetuate harm via incongruent targeted ads, echo chambers, or "rabbit hole" recommendations. Education around the machine learning (ML) behind Facebook (FB) can help users to point out algorithmic bias and harm, and advocate for themselves effectively when things go wrong. One algorithm that FB users interact with regularly is User-Based Collaborative Filtering (UB-CF) which provides the basis for ad recommendation. We contribute a novel research approach for teaching users about a commonly used algorithm in machine learning in real-world context – an instructive web application using real examples built from the user's *own* FB data on ad interests. The instruction also prompts users to reflect on their interactions with ML systems, specifically Facebook. In a between-subjects design, we tested both Data Science Novices and Experts on the efficacy of the UB-CF instruction. Taking care to highlight the voices of marginalized users, we use the application as a prompt for surfacing potential harms perpetuated by FB ad recommendations, and qualitatively analyze themes of harm and proposed solutions provided by users themselves. The instruction increased comprehension of UB-CF for both groups, and we show that comprehension is associated with mentioning the mechanisms of the algorithm more in advocacy statements, a crucial component of a successful argument. We provide recommendations for increased algorithmic transparency on social media and for including marginalized voices in the conversation of algorithmic harm that are of interest both to social media researchers and ML educators.

Introduction

Imagine a teenager who is struggling with body image. Every day when this teen logs into their Facebook, they see advertisements for beauty and diet products, images of bodies that don't match their own, and posts from friends discussing eating habits. If the teen has no idea that such content is algorithmically curated for them based on their behavior *as well as the behavior of their friends*, how will they ever advocate for a better online environment?

Widespread machine learning (ML) literacy belongs across all levels so that anyone can understand the algorithms they interact with, resist potential harms, and have a say in policy change. Numerous examples illustrate how ML algorithms driving FB advertisement recommendations can either directly or indirectly harm users. On Facebook, the news, posts, and suggested contacts a person sees are all driven by algorithms that make use of the underlying social connections (i.e. network) among users to infer and serve content of possible interest. Despite many positive experiences that result, there is a growing recognition that these systems also contribute to problematic, and pressing, social phenomena. Work on the growth and consequences of echo chambers, in which political and social opinions reflect and reinforce one's own, [390, 53] and studies of the perpetuation of gendered and racial biases through content recommendation [480, 361], for example, highlight the ways in which algorithmically driven

recommendations based on social network structure can lead to isolated and insular communities that reproduce harmful associations.

A classic algorithm used in ML that often powers recommendations is User-Based Collaborative Filtering (UB-CF), which ranks social contacts according to similarity to serve as a pool for potential recommendations. A key feature of UB-CF is that it relies on the connections among users to search for possible recommendations; in other words, data for ad recommendations can come from your friends and not just your personal characteristics and behaviors. Social media users, on average, do not realize that the content they may see is not only based on their own behavior, but also heavily influenced by what their friends are posting, liking, and clicking on. This process is hidden from the user, whose "likes" may leak into their network or whose social connections may leak into their personal recommendations without ever being notified. This might manifest as getting caught in a cycle of dieting or substance abuse ads, specific political values, conspiracy theories, or ads incongruent to your evolving gender identity.

There is potential for collective advocacy to draw attention to these issues of algorithmic harm via algorithmic resistance [485, 259]. The voice of the user provides us with insight about the ramifications of our AI systems. Therefore, it may be fruitful to explore how ML literacy efforts could work in tandem with self-advocacy skills – a concept primarily talked about in disability studies. Successful advocacy involves identifying the problem and articulating critiques and suggestions in order to adequately meet your needs [194, 498, 468]. A key step in advocating for one's needs is to understand basic properties of the potentially harmful system [397]. Previous work has demonstrated that using personal data may contribute to better advocacy arguments in ML contexts [397, 380, 267].

This work presents a novel research approach to not only teach social media users about algorithmic mechanisms and potential harm using their *own data* to demonstrate common algorithms underlying the platform's behavior, but also as a probe for user reflections on interactions with ML systems. This particular case study explores ad recommendation via UB-CF, a technique employed by Facebook Engineering to provide recommendations to more than a billion people across the globe. While FB modifies standard UB-CF, they likely keep the core idea of using similar users' tastes to identify candidate recommendations to serve to other users [257]. We designed a web app to teach these core ideas to participants as both instruction and as a prompt for surfacing potential harms.

In this intervention, not only do users get to see what kinds of data FB has collected on them, but they are able to trace how that data can be used by algorithms like UB-CF to generate new recommendations. Through investigating one's own data and learning core features of collaborative filtering, we predict that participants may feel empowered to manage the ads they see and be more vocal about what they want from social media companies. They may be motivated to unfollow certain accounts or be more careful about what they click on. They may even take active measures to click on content dissimilar to their interest in order to confuse or diversify what they see, a form of "gaming" the system once they understand the underlying mechanisms. They may even begin to think of their interests as something that can affect their whole network, leveraging the algorithm to promote societal change. We demonstrate how self-advocacy arguments describe these potential solutions post-intervention.

The web app we built teaches FB users (across all levels of Data Science experience) about UB-CF using their own personal Ad Interest data so that we can probe for not only accuracy of comprehension, but also for advocacy arguments about how UB-CF may potentially contribute to harm. While this case study uses the example of UB-CF, our approach can be extended to any common algorithm used by social media platforms. The main idea is to present instruction *in a natural and personal context* of how the user would encounter these algorithms in the real world. Therefore, while the findings in this case may be influenced by FB users' prior notions about and experience with Facebook, they represent a real-world intervention for promoting ML

literacy in an ecologically valid context.

Beyond simply learning about a particular algorithm, we care about the voices of users and how they articulate the potential harms caused by such algorithms in the context of FB recommendations. Because FB users are familiar with the platform and their own experiences, they can leverage this history to further understand UB-CF and potential ramifications of such an algorithm. We qualitatively extract themes from the user's advocacy arguments to look at how users express themselves after learning about UB-CF in the Facebook context. Because some of the most egregious harms tend to impact marginalized people specifically ([430, 366, 12, 23, 212], we particularly consider the voices of marginalized individuals, the kinds of harms they identify and possible solutions they suggest.

We formalize these aims with the following research questions:

> RQ1: Does this tool effectively teach how ad recommendation via UB-CF works on Facebook?
>
> RQ2: How do users advocate for themselves regarding potentially harmful ad recommender systems on Facebook after learning about the algorithmic mechanisms?

Foreshadowing the primary contributions of this work, we find that this interactive tutorial is a useful and novel approach to not only teaching social media users about ML algorithms in context, but also as a probe for harms, advocacy, and possible solutions. Through the case study, we provide details on how an ML tutorial can effectively integrate personal user data for instruction, allowing the user to relate more to the domain and use their own expertise to highlight specific concerns they may have about the algorithmic systems and platforms. The application as a probe encourages users to think about algorithmic harm and center *themselves* in this discussion, which can work well alongside researchers inferring user needs.

For this specific case study, we find that ML novices demonstrate high accuracy on a UB-CF comprehension task after participating in the tutorial, and that learning the basics of UB-CF increases the likelihood that the learner's advocacy argument will include mention of the network as a whole, as opposed to just their own behavior. This recognition opens new explanations of harms as well as solutions. We also surface themes of self-advocacy from marginalized users, who often provide personal examples of potential harms of UB-CF on Facebook; from accidental LGBTQ violence to the effects of pervasive dieting ads on individuals with eating disorders. Together these findings offer new directions for work on social media platforms, as well as studies of fairness and bias in these settings, by researchers, designers and policymakers.

Related Work

While research in ML literacy and user empowerment is relatively new, there are a few key spaces that need to be discussed in order to critically engage with the rest of the content in this paper. Current ML literacy efforts range from teaching ML concepts in K-12 [149, 234, 532] to critically deconstructing data and AI practices in the Science and Technology Studies space [50, 120, 115, 119, 387, 424]. One way to facilitate ML literacy is by using personal and relevant data, which has been explored by several research studies in recent years [267, 380, 38]. However, this work largely centers around formal instruction as opposed to general literacy for users of ML-based systems. Algorithmic systems continue to perpetuate oppression in the form of silencing, censorship, amplifying, and potentially harmful recommendations, particularly due to the structure and mechanisms of social networks [162, 94, 413]. Critical works often address these harms and their repercussions [13], but there is yet to be much research in the space of empowering novices to *overcome* these harms.

Model interpretability is one aspect of promoting general ML literacy, but explanations of model behavior are not enough to teach novices how ML will function in the future [445, 90]. Instead, involving the user in a familiar domain and probing for self-advocacy seems to have better results for general literacy [160]. The reason why self-advocacy is likely a successful frame for teaching is because when learners are prompted to reflect on their own experiences with a technology, they surface specific harms and must identify how algorithms perpetuated that effect. We know from learning sciences that using relevant and personal examples helps the learner connect to the material; in the case of self-advocacy prompts we are asking learners to not only surface problems of their own but also reason about them, enhancing the learners ability to connect and assemble knowledge. They draw upon the mechanisms of the model to reverse-engineer what is happening to them in their own context. This also allows us to study how marginalized communities are affected differently by the same algorithm.

Especially when studying anyone in a marginalized population, we need to understand that algorithms have differential effects [64, 153, 252], from Facebook [79] to health technology [362]. An app that works perfectly fine for an abled body could be destructive to someone who needs accessibility accommodations [466]. Data collection for those who fit within the gender binary may go unnoticed, where those outside the binary are consistently faced with rejection of their identity in tech[49, 269]. Racial biases in recommender systems, criminal justice, and health algorithms are increasingly part of the AI/ML research space [2, 179, 362, 50]. We know that trust from stakeholders matters [34, 523]; practitioners need to learn to design fairly and address the bias of their own systems [236]. This paper supports arguments that literacy from the ground up is an important way to facilitate change – stakeholders can speak up for themselves about how ML driven systems, such as social media platforms, are affecting them.

Designing a Web Application to Teach UB-CF with Personal Facebook Data

Facebook allows individual users to download a host of personal data, including Facebook's beliefs about the user's Ad Interests. While many algorithms underlie ad recommendation on Facebook, for this case study, we decided to design a tutorial to introduce users to one relatively simple (and commonly used) algorithm – User-based Collaborative Filtering – incorporating the user's own personal data in the instruction. We intend for this case study to be illustrative of the design approach and research probe, providing recommendations for other tutorial designs in the Discussion.

When a user downloads their personal data from Facebook they will find, contained in the files, a list of Ad Interests. The list contains words, products, people, media, and concepts that Facebook believes may be relevant to the user. (Ad Interests are not ranked in any way.) However, it is not explicitly clear how these Interests are curated; they are certainly not selected by the user for the purpose of personalizing advertisement. Our design demonstrated core principles of UB-CF on this data for a person with no prior experience with Data Science. The core principles of UB-CF were identified by synthesizing the simplest form of the algorithm:

1. Recommendations made to a user can be based on features of other similar users
2. A similarity metric is used to determine which users are the candidates to use for these recommendations
3. The most similar users' interests can be the recommended items

We built the web app using the Shiny package in R to administer the study intervention. The study procedures were reviewed and approved by our university's IRB. Recruited participants

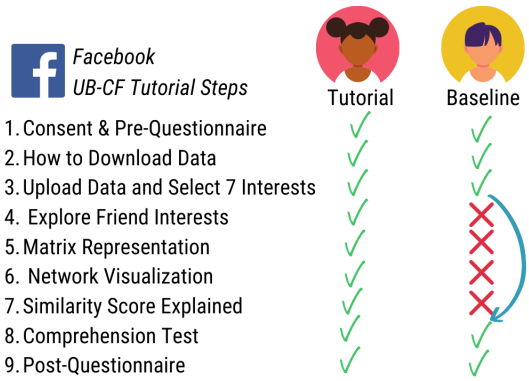


Figure 3.2: The steps that each group saw in their condition.

visit the domain and encounter on the first page a consent form; this is followed by an introduction with instructions on how to download your Facebook data. We specified that only Ad Interest data was necessary for this tutorial, and we asked users not to explore data prior to the study. Relying on manual data download and upload gives more agency, transparency, and privacy to our participants. We explicitly highlight the fact that no Facebook data would be recorded in the study and that our app only records survey responses. We do not observe the participants' Facebook data. Following the instructions, participants completed a baseline survey while waiting for their Facebook data to download.

In order to further personalize the experience, individuals could upload an image for their own avatar and enter their name. We did not store any personal data. For the purpose of illustrating the participant experience, imagine Shuri from *Black Panther* had a Facebook profile, shown in Figure 3.3b. In the Marvel Universe, Shuri is a Wakandan princess, scientist, and technologist responsible for much of the technological innovation of Wakanda.

Once a participant uploads their Ad Interests data it appears in a searchable, sortable, and paginated table (see Figure 3.3a). Interests were presented in random order to allow the participant to see a range of topics. The user was asked to select 7 Interests to use for the remainder of the tutorial. Next, the participant saw their personal avatar and three hypothetical friends, each with listed interests underneath their avatar (all taken from the original data provided by the participant to ensure some overlapping interests). The user could input names for each of the friends and customize their appearance if they desired, see Figure 3.4.

Imagine Shuri selected the following interests from her own data: *Combat sport, Community issues, Feminism, Hero, Science, Software,* and *Technology*. Figure 3.4 shows that Friend1 has two overlapping interests: *Combat sport* and *Hero*. Friend2 has an overlap of five interests: *Combat sport, Hero, Science, Software,* and *Technology*. Friend3 has an overlap of three interests. One of the friends is guaranteed to have the most overlap with the study participant (the pre-programmed size of the sample of overlapping interests is unique for each hypothetical friend) and therefore only one friend will be rated as the most similar to the user, avoiding complications with the UB-CF algorithm.

Next, participants saw Ad Interest data represented in matrix form. The columns represented the four people – in this case Shuri and her three friends – and the rows represented all of the possible Ad Interests among the four participants. Cells in the matrix are coded as 1 or 0 if the person in that column is interested in the interest on that row or not, respectively. Providing a matrix representation of the data (as one might use in ML model development) was used as an additional measure of comprehension and engagement with the tutorial. We were curious to see if ML novices who were engaging with their own personal data would be able to draw inferences from this data format. Interpreting the matrix is not trivial, and it is not immediately obvious

Show 10 entries Search:

	Interest
141	Recovery approach
142	Disney Princess
143	Social networking service
144	Oakley, Inc.
145	Thesis
146	Hero
147	Dogs
148	Psychoanalysis
149	Future (rapper)
150	The Economist

(a) Facebook interests data is visualized in a clickable, sortable, searchable table. Participants read that items appear in random order.

Shuri

Combat Sport
Community Issues
Feminism
Hero
Science
Software
Technology

(b) Selected interests are highlighted and appear in the list of selected interests.

Figure 3.3: Shuri selects 7 interests actually relevant to her.

which of the person pairs are most similar. In order to quantify which friend is the most similar to them, the participant would have to count the number of rows where both they and their friend had a 1. We asked participants to describe their process in determining which of their friends was the most similar to themselves. Participants were asked to select in a forced-choice response who they thought was the most similar to them. They received feedback on whether or not they were correct. Results from this probe are peripheral to the central arguments of this paper and are therefore not discussed at length, but it is interesting to note that 90% of all participants who saw this question got it correct (N = 35/39) and 87% of *not* Experienced participants got it correct (N=27/31), suggesting that ML novices in the study were able to understand the data presented as a matrix.

Next, participants saw a network visualization of their shared interests with their most similar friend, as seen in Figure 3.5. Figure 3.5 is a bipartite network, representing users and their interests as two types of nodes. Bipartite networks are a common way to represent users and interests in recommendation based systems, and are often the underlying data structure used in collaborative filtering algorithms. Users are linked or tied by their interests; the visualization thus allows shared interests (those shown in blue) to be easily identified. Interests that are held by their friend, but not themselves, are highlighted in green; these interests of their friend are likely to be recommended to the participant (differentiated with a green dashed arrow). The user can download their own network image as a data keepsake [297]. The participants saw a chart detailing the computed similarity metric for each pair of persons, sorted by similarity with the most similar pair at the top. Similarity was computed using cosine similarity [416], which results in a similarity measure between 0 and 1, with 1 meaning the two vectors are identical. The most similar friend, ranked most highly in this list, is the one whose interests are used as recommendations to the participant. Together, this exercise of exploring one's own interests, a friend's interests, and the overlap of interests provides the basis for a simple UB-CF model.

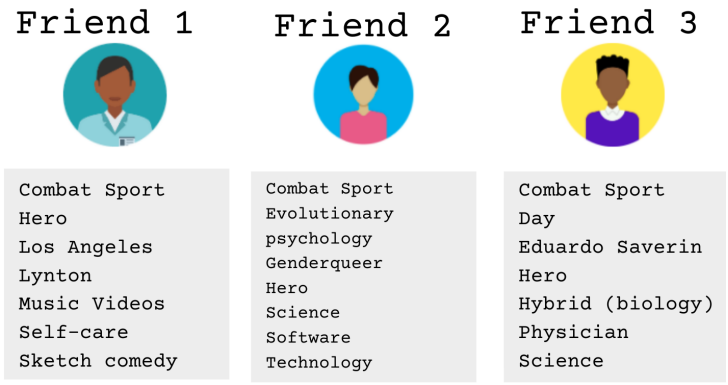


Figure 3.4: The application interface generates three hypothetical friends, each with some overlapping interests to the study participant. A single friend is guaranteed to have the most overlap. In this example, Friend2 shares the most in common with the participant, Shuri.

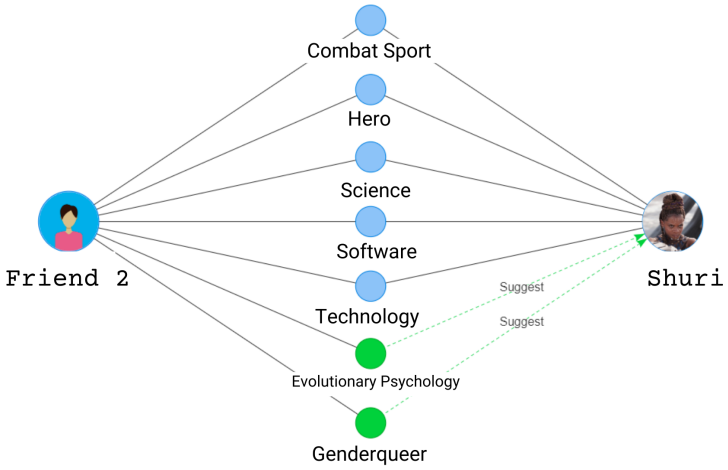


Figure 3.5: The participant saw a network visualization of shared interests between them and the most similar friend. Interests that may be recommended to Shuri are shown in green (*Evolutionary Psychology and Genderqueer*).

To test the efficacy of the instruction, and to establish a baseline, participants then took a comprehension test. They saw a New Friend who had 4 overlapping interests with the participants' original 7. Interests for the participant and the New Friend were presented side by side, as in Figure 3.6. Participants were asked to identify which interests would be recommended to them from the New Friend. They could select from any of the interests on the screen. The correct answer was the one-way "anti join" of the two lists (the Interests that appear on the New Friend's list but not their own). Following the comprehension test, a post-participation survey was the last component of the web application. The entire study took approximately 15 minutes to complete.

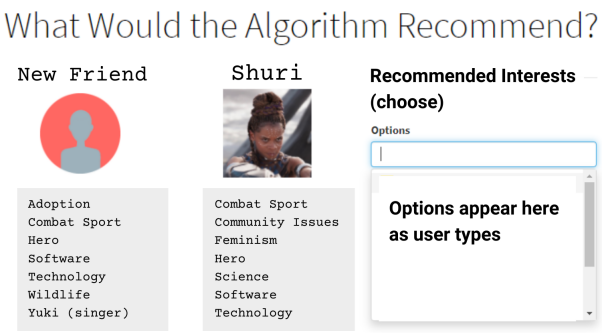


Figure 3.6: The comprehension test, asking them to report recommended interests based on data from a New Friend.

Methods

We recruited Facebook users ($N = 77$) of different Data Science Experience levels and tested their comprehension of UB-CF across two conditions: a Baseline Group with no UB-CF instruction (participants saw personal data) and a Tutorial Group where participants received UB-CF instruction on personal data. The Baseline Group included the data viewing page seen in Figure 3.3 and then proceeded straight to the comprehension task seen in Figure 3.6 to evaluate if simply *looking* at personal data would give users insight that Ad Interests may be estimated based on their friends. As part of the post-survey, participants responded to the prompt about potentially harmful recommendations on Facebook. These responses are analyzed to identify advocacy themes and to see how participants use new knowledge about UB-CF in their arguments.

3.3.1 Participant Recruitment, Demographics, and Limitations

We employed a quota-based sampling strategy to ensure participation along two axes: participants with a range of data science experience, as well as those in both marginalized and unmarginalized groups (self-identified), as seen in Table 9. We aimed for 15 members of each group, at a minimum. We recruited for this study through authors' social networks and via Facebook groups for various (non-academic) topics, some of them specifically for LGBTQ groups, disability groups, or activism groups. One challenge inherent in working with marginalized people is to establish trust in the researcher-participant relationship. This is difficult without the researcher disclosing their motivations and prior experience. Therefore, the first author relied on word-of-mouth recruitment and disclosure of their own marginalized identities. While this may produce bias towards specific characteristics the participant pool, it allows us access to users who normally would not participate in this kind of research. In fact, several marginalized participants indicated that they would not have done the study for another investigator. We recognize that marginalized people come in with biases, especially against Facebook as captured in our Pre-Survey, but emphasize that these voices are rarely represented in ML literacy research, which often relies on undergraduates or children. Future work is necessary to ensure a larger sample of marginalized individuals, and our results should be viewed with participant bias in mind.

83% of participants were between 21 and 36, with 9% between 37- 45 and 7% over 45. 86% of participants reported that they use Facebook every day, and 11% said they use it about once a week. The remaining 3% said they use it a couple times a month or less. Gender was not one of our independent variables of interest, though the write-in box for Marginalization status explicitly revealed at least 49% gender minorities in this sample. Other explicitly reported

Marginalized identities included race, autism, disability, women working in tech, religion, and sexuality. Some participants simply answered "Yes" or "No".

Participants used their own laptops to complete the study on their own time, ensuring an ecologically valid interaction with the tutorial. Participants were randomly assigned to the Baseline Group or the Tutorial Group as they visited the web application. Every participant received the same pre- and post- surveys; survey responses were stored in a mongodb database. Both Data Science Experience and Marginalization were measured in the survey, as described in Section "Pre- and Post-Surveys" presently. Participants volunteered their time in this study, and were not given any incentive, though several participants did it in order to see their own data (as reported anecdotally to the researchers). Future studies should consider compensation in order to enroll more marginalized individuals.

	Marginalized		**Unmarginalized**		
Data Science Experience	*Baseline*	*Tut*	*Baseline*	*Tut*	**Sum**
Experienced	14	4	5	3	26
Medium	2	7	4	3	16
Novice	10	8	6	11	35
Sum	26	19	15	17	77

Table 9: Participants by Data Science Experience and Marginalization status. Please note that χ^2 tests were not performed on any groups of < 5, but on aggregate groups depending on the research question.

3.3.2 Pre- and Post-Surveys

All participants completed a pre- and post-survey, which asked for basic demographics, willingness to share data, feelings of trust towards Facebook, and a free-response question about how they think Facebook generates recommendations for them. In order to identify participants' level of Data Science Experience, we asked: "Which of the following best describes your experience with Data Science, Computer Science, and/or Machine Learning?" Response options, shown in Table 10, were re-partitioned into Novice, Medium, and Experienced. While difficult to empirically verify, the partition is corroborated by accuracy on the comprehension task in the Baseline Group condition, where we would expect participants coming in with more experience to do better than those without experience.

☐ I don't know anything about any of those topics.	Novice
☐ I have a vague idea of how some of those things work, but with no formal instruction.	Novice
☐ I have taken classes in any of those subjects.	Medium
☐ I know a fair amount about those topics.	Experienced
☐ My job is in Data Science, Computer Science, and/or Machine Learning (e.g. I have the title of Data Scientist, or do Machine Learning work regularly)	Experienced

Table 10: Survey responses for Data Science Experience, along with partition used in analysis.

We asked participants if they identified as marginalized. Participants responded in detail, ranging across sexuality, race, gender identity, neurodivergence, immigration status, and dis-

ability. The first author re-coded the responses as a binary variable for anonymity. The survey asked:

> Do you consider yourself a part of a marginalized group? For example, this researcher is nonbinary. Your answer will NOT be shared or linked to your identity. Please describe below. We really appreciate your vulnerability. You may also choose to write "Prefer not to say".

We wanted to know if simply *looking* at your own Facebook data (Baseline Group) affected willingness to share data, and if learning about UB-CF affected that willingness. We asked: "Would you be willing the share the data you just downloaded of what Facebook thinks you're interested in if it were anonymized? Please check all options you would be comfortable sharing anonymized data with." Participants could select multiple choices from: University Researchers, Company Marketing Teams, Other Apps in Your Phone, Political Campaigns , Government Organizations, "I would not be willing to share it", or Other. 13% of participants changed their answers for whether or not they would be willing to share their anonymized Facebook data after looking at their data, and there was no distinct pattern linked to any of the relevant factors (Experience, Condition, Accuracy, Marginalized). Further investigation is needed.

The pre-survey also asked: "Do you trust that Facebook cares about its users and acts with their interests in mind? " following Lankton and McKnight. Only 1 participant said Yes. We also asked: "Do you trust that Facebook's algorithms have the ability to recommend things to you that you actually like?" and 37 participants said Maybe, 16 said No and 24 said Yes. We include these peripheral findings to establish a baseline for our sample and its biases, as well as to offer directions for future work.

In both the pre- and post-survey, we asked for a free-response to the following prompt:

> How do you think Facebook comes up with the list of topics that it thinks you might be interested in? Brainstorm as many ideas as you can. e.g. they gather data from what you click on

We then prompted participants for an advocacy argument:

> Imagine Facebook recommended something harmful to you. Use this space to describe what you think went wrong and what can be done about it.

3.3.3 Evaluating and Comparing Comprehension of UB-CF

We measure comprehension of UB-CF by asking participants which of a list of Interests from a similar *Friend* would be (algorithmically) recommended to them (see Figure 3.6). The correct answer is any listed Interest of the friend, that the participant does not currently have. We evaluated both the Baseline Group and Tutorial Group for their accuracy on this question. Importantly, the correct answer is a set of 3 Interests. As such, participants could give a partially correct response or a mix of both correct and incorrect responses. To quantify accuracy, we used an F1 Score to capture both Precision and Recall [417]. A perfect F1 score is 1. We compare F1 scores between both the Baseline Group and Tutorial Group, and across Experience levels to assess differences in comprehension. We hypothesize for RQ1:

> H1: Those with more Data Science Experience will perform better in the Baseline than Novices, but the Tutorial will improve accuracy for all levels of experience.

An important learning objective was to show to learners how recommendations come from not only their own behavior, but from the behavior of the friends in their network as well. To evaluate this more nuanced signal of UB-CF comprehension we use the textual responses from the survey question on how Facebook generates recommendations. Two coders analyzed the arguments for whether or not they mentioned friends, with an inter-rater reliability (IRR) of .96.

For example, one participant responded: *"The algorithm may have wrongfully computed a suggested interest for me based on interests of those I am **"friends"** with but do not share enough commonalities with."* We further hypothesize for RQ1:

> H2a: More participants will mention the effects of friends on recommendations following the Tutorial.
>
> H2b: Participants with more Data Science Experience may already know about the effects of friends on recommendations and will mention friend effects in both the Tutorial Group and Baseline Group.

3.3.4 Thematic Analysis of UB-CF Learning Objectives and Potential Harms on Facebook

We are interested in the kinds of potential harms that participants surface after learning about UB-CF on their own data, and how they advocate for themselves when prompted. To extract this from free-response advocacy arguments, we use thematic analysis, following the guidelines of [358]. Researchers developed an initial code set by reading a sample of arguments (without knowing the condition, marginalization status, or experience level of the respondent) and identifying recurring words and topics. Next, individual arguments were sorted into affinity groups of the same code, such as "LGBTQ experience" or "politics". At this point, some codes were collapsed in order to produce the final code set shown in Tables 11, 12 and 13. Each theme consisted of a group of arguments with a linking relationship between them that surfaced as a central description for that grouping; we tried to minimize the number of possible groupings to avoid overspecificity. After a period of open coding, two coders labeled the advocacy arguments with their themes, with an inter-rater reliability of .85. In our analysis, we focus on themes with several examples in the data, and discuss these findings.

Results

3.3.5 Participant Comprehension of UB-CF

This study aims first, as stated in RQ1, to evaluate whether participants demonstrate increased comprehension of UB-CF after learning UB-CF with personal data. We evaluate this aim by prompting participants for a fixed choice response with a correct answer, as well as in a free-response question asking participants to describe how they think UB-CF works. Observed F1 scores on the comprehension task, by Data Science Experience and Condition, are shown in Figure 3.7. As hypothesized (H1), participants in the Baseline Group group with more Data Science experience were more accurate on this task than Novices, with the Medium group in between. This result lends validity to the stratification of participants in terms of Data Science experience at baseline. Following the Tutorial, all levels of Data Science Experience improved to a statistically indistinguishable accuracy ($p = .54$), with a Kruskal Wallis test revealing a significant difference between the Tutorial Group and Baseline Group ($chi^2 = 20.061, df = 1, p < .00001$). To determine if this difference is driven by one level of experience or the other, we also conducted separate Mann Whitney tests that reveal differences between the Tutorial Group

and Baseline Group for Medium ($W = 11.5, p = 0.030$) and Novice ($W = 45.5, p < .00001$) levels of experience.

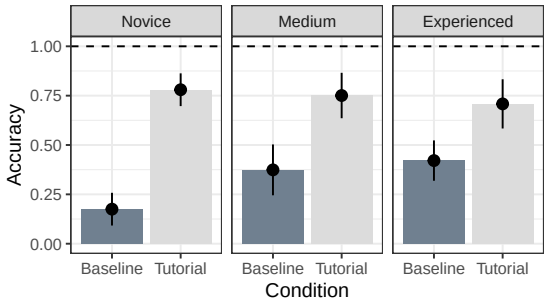


Figure 3.7: F1 score by Condition and Experience

Testing Hypothesis 2a and 2b, we find that before seeing their own data, 40% of participants included 'friends' in their answers in both the Baseline Group and Tutorial Group, indicating no difference in baseline knowledge before the Tutorial actually occurred. After the completion of the Baseline Group condition, the proportion mentioning friends increases to 60%. In the Tutorial condition, 83% of participants mention friends as part of how Facebook comes up with recommendations. This Tutorial effect is significantly greater than the Baseline ($\chi^2 = 3.9452, df = 1, p = 0.04701$).

3.3.6 Harmful Recommendations and Advocacy

Next, we address the question of how study participants advocate in the face of potentially harmful algorithmic recommendations, as stated in RQ2. We are interested in differences in advocacy arguments between the Tutorial Group and Baseline Group, and across Marginalization status.

3.3.6.1 Mentioning their Friend Network in their Arguments

We stratify participants, grouping those with High accuracy ($F1 >= .6$) on the comprehension task and those with Low accuracy ($F1 < .6$), using the distributions of F1 scores to determine this cutoff. Figure 3.8 shows that if you understand UB-CF (High Accuracy regardless of whether you learned from the tutorial or if you came in with the knowledge already), you are more likely to use that understanding in your advocacy argument. Recall, mentioning the innerworkings of a problem with actionable points is part of a successful self-advocacy argument [194]. We demonstrate here that if you know about the details of UB-CF, you will use those details to advocate.

Participants with High Accuracy mentioned the effect of friends in their advocacy arguments significantly more than those with Low Accuracy ($\chi^2 = 4.35, p = .037$). Figure 3.8 shows the proportion who mentioned "friends" in their advocacy, such as this iconic argument from the data:

*"A **friend** with other interests in common is into some weird shit and FB assumed that I'm probably into the same weird shit."*

3.3.6.2 Themes of Advocacy About Risks of UB-CF on Facebook

We also surface a non-exhaustive list of themes from participants' advocacy arguments when imagining that Facebook recommended something harmful to them. One category of themes

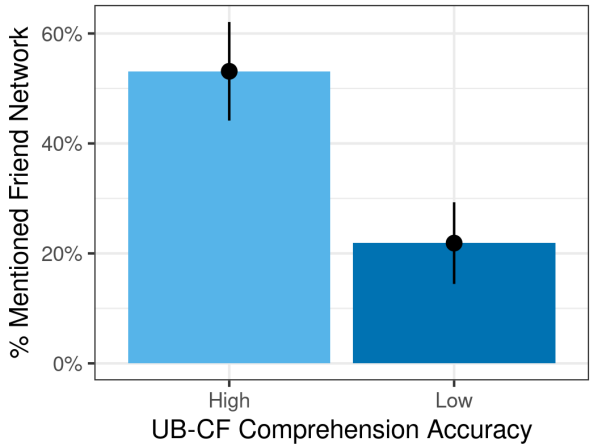


Figure 3.8: Those with High Accuracy in the UB-CF comprehension task mentioned the effect of friends in their advocacy arguments significantly more than those with Low Accuracy ($\chi^2 = 5.4, p = 0.020$)

that emerged was specific examples of harm from the influence of peers in the social network, examples seen in Table 11. Next, we see more general commentary about potential harms, generalizable to several behaviors or identities (see Table 12). We also uncovered some themes about solutions to the problem (see Table 13), which is a key part of successful self-advocacy.

Theme	Example Response
Eating Disorders	"Probably recommending diets to someone who has a history of eating disorders. I keep trying to hide those ads and mark them as "sensitive topic". Someone looking up a bunch of diets online is probably interested in diets, but recommending more diets might actually be harmful."
LGBTQ Experience	"Some of my friends/family are still extremely religious. If there's not a way to see if a recommended interest does not work with my current interests, I, a queer person (who fb knows is queer) could get like....recommended conversion therapy because of my conservative family."
Political Ads	"It seems to me that Facebook is factoring the opinions of your Facebook friends into the process. I don't think they should do this at all. People can be friends and not share opinions, especially on Facebook. A liberal person could be friends with a family member who supports Donald Trump. That doesn't mean they support Donald Trump."

Table 11: Themes of specific examples of harm from the influence of network peers.

Each of the themes in Table 11 may provide a basis for further research. Studying the effects of diet and beauty ads on social media users is key to understanding mental health and digital "hygeine" may reveal concerns specific to this context, such as the influence of visual content. The LGBTQ online experience is complex, with the internet often serving as a safe-haven for community but also perpetuating harm (e.g. enforcing strict gender roles or perpetuating violence). Understanding how non-experts perceive political targeting is vital to designing more transparent systems.

Table 12 reveals social media phenomena that can only be understand by representing user voices and what they notice, need, and are concerned about. Current work in this domain does not always capture phenomena that are obvious to actual users. We also surface potential

Theme	Example Response
"Hate following"	"Hate following happens across industries. Facebook is likely recommending something based on something the user hate followed instead of followed because they actually liked the topic."
Mis-information	"As I'm apparently put in a category of people caring about the environment, I get spammed with all things "natural", so a lot of scam and potentially dangerous "cures". I'd like to see ads making unsubstantial claims gone. Greenwashing should also be forbidden."

Table 12: Themes of algorithmic harm generalizable to many topics, focused on misinterpretation or misrepresentation of network relationships.

Theme	Example Response
Personal Behavior Change	"The algorithm took information from my engagement with something similar and used it to make a suggestion it believed would illicit future engagement. Selecting the "see less like this" option or intentionally not engaging in similar content in the future could help correct it. "
Suggestions to FB	"A friend showed interest in (or accidentally clicked or searched for information on) something harmful and it was then recommended to me. Doesn't sound like anything went wrong if it was designed this way. Sounds like it needs to be re-designed to prioritize an ethical process that sees users as human rather than passive money-makers."

Table 13: Themes about possible solutions to the problem

solutions offered in the advocacy arguments, in Table 13, not only to demonstrate potential solutions to the research community but to show that non-experts can be a valuable resource for design solutions.

Discussion

"*I would assume that someone else liked or had an interest in that thing, because people can hold harmful values without knowing it.. then those harmful ideas are spread around by Facebook's algorithms and if they go unchecked that can be really problematic. I think more transparency is good, too.*" – P31 (Novice after Tutorial)

In order to increase democratic participation in the design and use of algorithmic-based interactions on social media platforms, we first need to describe what Data Science novices already understand about these systems, and then thoughtfully develop ways for them to learn more about where, how, and why algorithms affect them in their daily lives. It is crucial that we provide pathways for self-advocacy for users to express their needs and potential or experienced harms as they engage with these systems.

This paper contributes: 1) insight into baseline knowledge of a sample of Facebook users about algorithms that underlay recommender systems across all levels of Data Science experience, 2) a successful and personalized intervention design to offer instruction to users about how algorithms work, and 3) empirical demonstration of an association between learning algorithm details and improving self-advocacy arguments when describing potential harms from Facebook's ad

recommendation.

At baseline, we find that the majority of our participants do not assume that Facebook uses their friends' data to inform ad recommendations. We see a general distrust in Facebook among participants, but a willingness to explore their own data and provide critique of Facebook's algorithm when prompted. We detail an intervention for teaching users about UB-CF on their own Facebook data, and compare the effects of simply *looking* at your own data to being instructed about UB-CF on that data in terms of comprehension and advocacy. We find that instruction improved comprehension for all levels of Data Science Experience, a boon for both ML pedagogy and widespread ML literacy. We empirically demonstrate an association between accuracy in UB-CF comprehension and using the mechanisms of the algorithm in an advocacy argument about potential harms from the system, providing synchrony between our quantitative and qualitative findings. Successful advocacy often involves negotiation, providing alternative solutions, and understanding how you are being harmed. We show that participants who successfully learned about UB-CF were more likely to use that knowledge in their arguments. Participants advocated for potential solutions, such as unfollowing or hiding ads, exposing themselves to content outside their usual interests to avoid echo chambers, demanding more transparency from Facebook, asking for a feature where they could filter *which* friends are used in recommendation algorithms, and leaving Facebook altogether.

Marginalized users used instruction on UB-CF to help express specific ways that algorithms contribute to harm on Facebook. Surfacing these hypothetical and experienced harms is vital if we – social media researchers and designers – want to meet the needs of marginalized stakeholders and understand differential effects of algorithms in these settings. The insight from Marginalized participants is valuable and detailed, ranging from concerns about religious influence on their queer identity (e.g. *"for instance, I am LGBTQ (closeted) and have friends who are more conservative/fundamentalist Christian. They may have interests that are harmful to my identity (e.g. pray the gay away) that would be pushed to me."*), to pointing out harms perpetuated by the beauty and diet industry (e.g. *"recommending diets to someone who has a history of eating disorders. I keep trying to hide those ads and mark them as "sensitive topic". Someone looking up a bunch of diets online is probably interested in diets, but recommending more diets might actually be harmful."*)

Pathways for Future Work

The results of this study are a starting point for future efforts both on ML literacy, as well as successful advocacy in cases of algorithmic harm. Here we suggest some specific pathways that we believe will yield important research and design; we aim to pursue many of these questions.

3.3.7 Machine Learning Education

Many machine learning lessons rely on unrelatable datasets such as the classic `iris` or `mtcars` datasets, or outdated and implicitly biased, the `Boston housing dataset` from the 1970s. Use of such datasets is perhaps unsurprising – they are clean and nicely demonstrate ML algorithms and models. Our work demonstrates one way to integrate relevant, interesting data into ML instruction. Because learners have expertise about their own experiences, they can apply this knowledge to ask targeted questions about algorithms and data they are learning about in context. ML education and literacy efforts, as well as research on these topics, should leverage relevant contexts and associated data such as social media algorithms, Google searches, face filters, and other ML systems that are being used in the real world. Doing so can increase comprehension and support learners to evaluate the content they see. See the section on Designing Tutorials with Personal Data for further guidance.

3.3.8 Self-Advocacy of Novice ML Users

Work on algorithmic bias and harm often disregards the knowledge and domain expertise of ML Novices, referring to them as "laymen" or "everyday people". Our results suggest that Novices can not only learn about ML topics, but also that they will use what they learn in advocacy arguments. Social media researchers have the ability to probe and highlight these valuable insights. The path to ML literacy involves both *top-down* (designers and engineers programming the systems) and *bottom-up* (users and stakeholders critiquing the systems) approaches. Platforms should provide more opportunities for user feedback specifically regarding the algorithmic portions of the user experience. This could mean offering more user *control* [84] such as sliders, filters, misinformation flags, or access to the "weights" on their newsfeed content, for example. Researchers should take care to include perspectives of users when discussing algorithmic harm, especially because what researchers assume is harmful may differ from what users feel is harmful. Algorithmic bias can be especially damaging for marginalized communities. Researchers have an opportunity to focus on the researcher-participant relationship in order to successfully involve marginalized users in a non-extractive way. For example, see the frameworks presented in *Data Feminism* [116].

3.3.9 Designing Tutorials with Personal Data as Research Probes

For future design of ML tutorials using personal data, here are a few guidelines to consider based on our experience and research agenda in this space.

3.3.9.1 Considerations for Using Personal Data

The user must consent and know if their data will be stored and how it will be used in any research study. It is also possible, as illustrated here, to make use of personal data while also maintaining participant privacy. For this work, we allowed the user to upload their own data files and did not store their personal data, only their responses to our prompts. Personal data should never be used in a group setting. Any instruction should be mindful of potentially sensitive content that the user may not expect or intend to inspect. This could be explicitly sensitive content such as aspects of identity, or even less obvious content such as recent experiences or events that are recorded in the personal data (e.g. reminds the person of a breakup or death). We recommend that participants filter their own data for what they would like to use for the remainder of the tutorial (as we did in this study); while it is difficult to avoid unanticipated sensitive data altogether, this process can lessen exposure and harm.

3.3.9.2 Considerations and Ideas for Designing ML Tutorials

Our strategy was first to identify platforms that allow the user to download their own data (such as Facebook, Instagram, Google, Twitter, etc.), along with the data formats available. Data is, in fact, often a starting point for research of this nature. One might next consider common ML algorithms used on the focal platform. Social media systems are build with layers of algorithms, and as such there are likely many to consider as topics for instruction, e.g. recommender systems, misinformation or hate speech detection, image recognition, etc. We found it key to consider which algorithms have a significant impact on the user, not only in terms of frequency of interaction but also potential for harm. Algorithms that users may not even realize are powering their experience can be especially fruitful targets [393]. A final consideration is the instructional demonstration itself. It needs to realistically incorporate personal data to show the user a basic form of the algorithm; visualizations and other tools can be helpful in communicating about algorithms to participants.

3.3.9.3 Examples for Future Tutorials

In light of the above considerations, and based on our experiences in this project, we believe the following interventions will be promising directions for work:

- demonstrating image classification or object identification on a user's own images, especially with regards to images that get banned or reported
- introducing participants to algorithms powering *DeepFakes* using their own shared videos, and reflect on the dangers of such tools
- show clustering algorithms on a user's personal network on social media platforms

Limitations

Given the sample size and recruitment strategy in this work, results should be replicated on a generalizable sample to confirm the efficacy of this tutorial. This work represents a roadmap for future efforts, including a novel approach to include the user's own data in learning about algorithmic systems and potential harm. A notable limitation in this work was the short responses given for advocacy arguments; making them less fit for in-depth qualitative analysis – necessary to fully understand stakeholder voices. Another limitation is the lack of random sampling of participants; we instead purposely highlight the voices of marginalized groups as a first step in exploring this topic. We did not imitate Facebook's exact algorithm in our instruction, but instead used a rudimentary version. While this may affect the real-world applicability of user advocacy arguments, we believe that core literacies should transfer to other similarity-based recommendation algorithms. Future work might attempt to replicate results across different algorithms, or alternatively, build from real cases of harm before prompting advocacy.

Conclusion

Social media users interact with potentially harmful algorithms every day, but through education, users are able advocate for themselves when such harms occur. This paper represents a step to democratizing ML by promoting understanding and advocacy for Facebook users. Whether its Shuri from Wakanda or a teen scrolling social media, we can all benefit from increased literacy and self-advocacy skills to participate in an ML-driven world.

Verbatim Text

Verbatim text ends here.

3.4 Chapter Summary and Contributions

✓ Summary

- ✓ Corporations, government organizations, and educational institutions have all put forth increasingly substantial efforts to reduce bias and discrimination in AI as the call for Responsible AI amplifies. This chapter summarises some of the major positions on AI Ethics Standards across industries – comparing the approaches of 'Big Tech' like Google, Amazon, Meta, and Microsoft with government initiatives and statements from the Association for Computing Machinery (ACM).
- ✓ While there are several initiatives for Responsible AI, there is also evidence to suggest that AI ethics guidelines are not immediately practical [174, 336]. For example, programmers given the ACM Code of Ethics had no significant difference in decision making from a control group [314]. AI ethics efforts have been widely criticized for "lacking teeth" [403].
- ✓ In addition to summarising AI Ethics Guidelines and Responsible AI efforts, I review the current landscape of AI Ethics education. I demonstrate that AI Ethics education typically centers around only a few high profile cases, and is not embedded or integrated into the technical material.
- ✓ I provide recommendations for effective AI ethics education, distilled from the current literature. The recommendations are that AI ethics education ought to be: *specific, prescriptivist, action-centered, relatable, empathetic, contextual, expansive, preventative* and *integrated*; elaborated on in Table 8.
- ✓ I provide a published work on one example application for embedding ethical education into a technical machine learning lesson on recommender systems – using the technical details to highlight potential for algorithmic harm, with users providing additional feedback on how the algorithms might recommend something harmful to them [398]. This case study simply demonstrates one avenue for integrating ethical thinking into technical AI education that centers marginalized individuals' concerns.

△ Data Science Tip

Register and Spiro [398] provide an example of incorporating personal data, relatable experiences of everyday algorithms, and probing for algorithmic harms. All together, students can engage in a pipeline of Data Science that considers policy and data access, the actual data source, the technical details of the model, the social context of the system, and the risks to the end-user. Altogether, this is a more complete walkthrough of the entire process, contributing to better Data Science skills overall.

★ Contributions

- a distilled summary of current AI Ethics Standards and the various ways they fall short
- an overview of current AI Ethics education approaches, and my recommendations on effective AI Ethics education. The recommendations are that AI ethics education ought to be: *specific, prescriptivist, action-centered, relatable, empathetic, contextual, expansive, preventative* and *integrated*; elaborated on in Table 8.
- published work demonstrating an effective application of embedding AI ethics into technical instruction of collaborative filtering, where learners explored algorithmic harm from recommender systems on Facebook using their own ad data.

Part 4
Fostering Care and Empathy in AI Education

"Remember to imagine and craft the worlds you cannot live without, just as you dismantle the ones you cannot live within"

– Ruha Benjamin

Abstract

I demonstrate through preliminary empirical research that students who relate to scenarios of algorithmic harm find them more urgent, and by including a wide spectrum of algorithmic harm scenarios in an AI curriculum we can broaden participation and foster empathy. My findings also show the need for more pedagogical support for practical decision-making skills in AI contexts.

Chapter 3 introduced the notion that AI ethics standards may not be sufficient to adequately prepare Data Scientists to make ethical decisions in practice. My research thus far has indicated the utility of using personal data and lived experiences to connect to cases of algorithmic harm, specifically in a technical setting. In both [397] and [398] my findings demonstrate that Data Science students who relate to the topics of algorithmic harm will use technical mechanisms of those machine learning algorithms in advocacy arguments to point out both its social impacts. My work on algorithmic precarity [400] and discriminatory content moderation [401] exposed nuanced cases of algorithmic harm that may be more *expansive* than the typical scenarios presented in AI ethics courses, and are therefore fruitful ground to appeal to multiple perspectives in an AI classroom. I argue that successful AI ethics education is one that helps to retain underrepresented students, and fosters empathy in students who may not relate as strongly to cases of algorithmic harm. It is also ones that gives students practical action to take to repair models and mitigate harm, as well as preparation for potentially uncomfortable decision-making in the workplace. This chapter reports the results of a preliminary study on student reactions to 30 real-world cases of algorithmic harm, measuring a variety of variables to operationalize *care*. Students rate each of the 30 scenarios (many taken from Table 5 in Chapter 2) on how personally relatable they are, how urgent the scenario is, how good of a fit they personally would be to fix the problem, and whether or not they should delay or deploy the flawed software. We find that many scenarios show a positive correlation between relatability, urgency, and personal good-fit. We also find evidence of demographic affinity towards scenarios with algorithms that directly impact one's gender or race. However, we see an overreliance of "life-or-death" reasoning, with students using the potential for death as their main ethical decision making tool, with difficulty to think in more nuanced ways. We provide sample curricula and recommendations for teaching AI ethics using these empirical findings, and promote an ethics of care (EoC) approach to support inclusivity and foster empathy in the AI classroom.

4.1 Author's Motivation

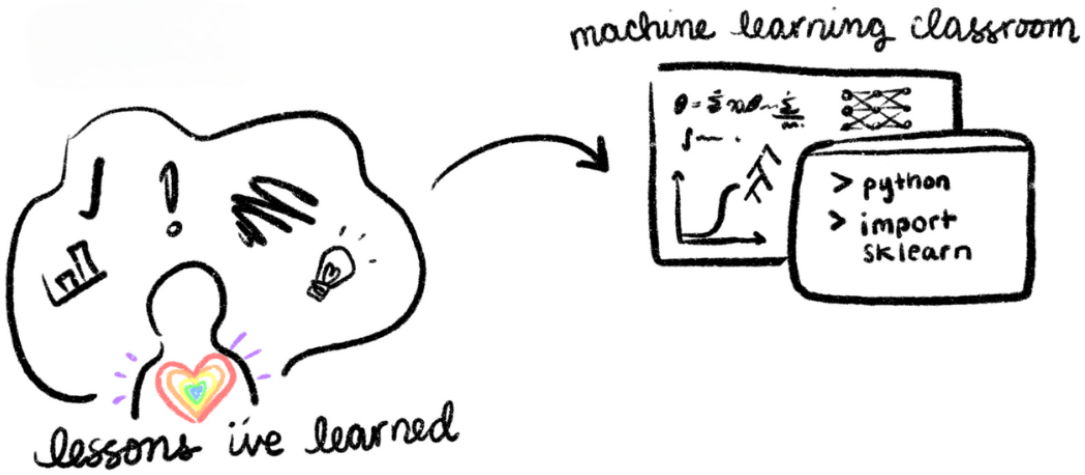


Figure 4.1: Original artwork of a cartoon person with a rainbow heart and the words "lessons I've learned" pointing to Data Science icons and the words "machine learning classroom", implying all the work that has come before will now be integrated into final recommendations on how to effectively teach AI.

One of the most common questions I get asked about AI ethics is *"What should we teach?"* If you've followed along so far, you know that teaching AI ethics is more important than ever, in order to mitigate algorithmic harms and to work towards a sustainable future. Teachers for as young as elementary school are adopting AI curricula that include bias and fairness concepts. There are increasing calls for integration of AI ethics throughout Data Science programs, and more demand from companies to practice responsible AI. As the world grapples with applications of generative AI, policymakers and government are wrestling with the appropriate checks and balances on AI technology. In a landmark resolution from the UN General Assembly in 2024, the US Ambassador to the United Nations Linda Thomas-Greenfield said:

> *"So let us reaffirm that AI will be created and deployed through the lens of humanity and dignity, safety and security, human rights and fundamental freedoms. Let us commit to closing the digital gap within and between nations and using this technology to advance shared priorities around sustainable development."*

I demonstrate in Chapter 3 that we must break down statements such as these into actionable, practical, and context-specific steps. Creating and deploying our technology through the lens of "humanity" and "dignity" will mean very different things in different contexts and to different people. "Shared priorities" and "sustainable development" will also mean different things. It is valuable to see such high-level representatives speaking on AI safety, and it is partially our responsibility as educators to translate these expectations into actionable steps for the Data Scientists we are training. Given the demand for safe, human-centered, responsible, fair, and ethical AI: how do we effectively prepare future Data Scientists to think in compassionate, ethical, and responsible ways? *What do we teach and how do we teach it?*

My research thus far has shown the importance of *integrating* AI ethics concepts into technical lessons. Table 8 in Chapter 3 outlined the characteristics of effective AI ethics education: *specific, prescriptivist, action-centered, relatable, empathetic, contextual, expansive, preventative and integrated*. This chapter explores empirical work in support of those recommendations,

reporting the results of a pilot study exploring how students responded to various cases of algorithmic harm[4]. I am deeply curious about what motivates learners to genuinely engage with topics of AI bias and algorithmic harm, so that we might build a better future together. The future of AI that I envision is compassionate, careful, accountable, and kind.

Our world will continue to grapple with issues of injustice, social inequities, and human rights concerns. It is not only Data Scientists and computing professionals who lack skills in these areas – these are *wicked problems*, affecting every industry and continually compounded by constraints on resources to solve them.

Definition 4.1

Wicked problem: a social or cultural problem that's difficult or impossible to solve because of its complex and interconnected nature. They are difficult to formally define, lack clear 'stopping points' of when they are solved, can be described as symptoms of other problems (interconnected), and each trial solution has cascading effects. Examples of wicked problems: *poverty, climate change, education, homelessness, sustainability* [404].

As previous chapters have discussed, it may be the case that those in the computing field are particularly unprepared to confront social problems. This is *not* due to the common misconception that 'computer geeks don't have social skills'. Instead, it is likely due to both normative assumptions about computing work itself, and the lack of preparation given to learners in the industry. In a field like social work, public health, or education, one can expect to be working with people – issues of social inequality are often core to educational programs and practice. For the computing industry, there are pervasive assumptions of positivism and mathematical operationalization of human problems – a detachment from the lived realities of those affected by computing systems [116]. Computer Science students are trained in riddles and formal solutions, with clever "a-ha!" moments to solve the puzzle. Unfortunately, wicked problems cannot be solved with clever tricks – and must be instead approached as a multi-tiered collaborative process. Knapp [271] states:

> *"The first is to shift the goal of action on significant problems from 'solution' to 'intervention.' Instead of seeking the answer that totally eliminates a problem, one should recognize that actions occur in an ongoing process, and further actions will always be needed."* [271]

So, while it is not unique to those in the computing field, Data Scientists-in-training particularly need a shift from solutions-oriented approaches to human-centered and iterative *interventions* that keep a wide array of stakeholders in mind. Research suggests that: a) Data Scientists suffer from 'in-group bias' and care more about problems they can imagine actually happening to them [208, 209]; b) curricula lack scaffolding and support to engage in what seem like controversial topics of race, gender, and politics [395]; and c) AI ethics considerations are often self-contained at the end of curricula, furthering their position as something to be 'othered' and potentially ignored [180]. Chapter 3 explored theoretical work regarding potential interventions for practical ethics in AI, which I synthesized into recommendations on characteristics of effective AI ethics education. However, despite the urgent calls for more responsible AI practices, there are few empirical studies in this area [533]. In the following pilot study, I gather baseline information about how Data Scientists-in-training perceive a variety of algorithmic harm scenarios. I test their in-group bias in terms of how their personal relatability to a problem impacts its perceived urgency. I also look at ethical frameworks that students use to make decisions about deployment

[4]many of which were taken from Table 5 in Chapter 2

of various AI technologies, outlining the various pressures and cognitive processes impacting students' willingness to delay deployment on biased software. These results allow us to propose a curriculum design that is both inclusive and practical.

The following are preliminary pilot results that have not yet been peer reviewed. Special thanks to my collaborators Dan Schneider, Emma S. Spiro, and Nicholas Weber.

4.2 Empirical Evaluation of Student Care for 30 Algorithmic Harm Scenarios

Introduction

With the rising demand for AI technologies, have been numerous examples of algorithmic harm and unintended consequences – spanning discrimination [80], lack of equal opportunity [360], medical errors [362], safety concerns[13], exploitative AI [12], and retraumatizing effects of algorithms [98], just to name a few. Chapter 2 provides an extensive list of algorithmic harm examples in Table 5. In recent years we have also seen growing interest in and calls for more integrated FATE topics: fair, accountable, transparent, ethical AI [42]. Sometimes referred to as 'Responsible AI' or 'AI Safety', there is growing concern over the ethical considerations of the AI technology we create. One point of intervention is in AI education itself, preparing future Data Scientists to engage in the technical aspects of AI but also its social impact.

Definition 4.2

Ethics of Care (Care Ethics): A feminist ethical theory tracing back to Carol Gilligan [187] that posits moral action centered on interpersonal relationships, social contexts of a moral dilemma, and empathy for the most vulnerable or dependent. [442, 154, 15, 354, 353]

In Chapter 3 I covered the 'toothlessness' of AI ethics [336], referring to the phenomena that AI Ethics guidelines do not translate into practical action, nor are they enforced in meaningful ways. For example, one empirical study demonstrated that the ACM Code of Ethics had *no effect* on programmer decision making in their software design compared to a control [314]. So what does impact ethical decision making, if not ethical guidelines? What factors impact how much a student cares? One thing that has been called for by various scholars is the integration of more ethics of care, also known as care ethics, in order to foster relatability to cases of algorithmic harm [395, 100, 209]. For example, Hagendorff [209] define care in the context of AI in *"AI virtues – the missing link in putting AI ethics into practice"*. They write:

> "**Care** means to develop a sense for others' needs and the will to address them. Care has a strong connection to empathy, which is the precondition for taking the perspective of others and understanding their feelings and experiences. This way, care and empathy facilitate prosocial behavior and, on the other hand, discourage individuals from doing harm. In AI ethics, care builds the bedrock for motivating professionals to avoid AI applications from causing direct or indirect harm, ensuring safety, security, but also privacy preserving techniques. Moreover, care can motivate AI practitioners to design AI applications in a way that they foster sustainability, solidarity, social cohesion, common good, peace, freedom and the like. Care can be seen as being the driving force of the Beneficial AI movement."

For a simple solution we ask: "What algorithmic harm scenarios do students care about?" We could take those scenarios and include them in our curricula so that students would be more likely to engage with the material. In a survey design, you can imagine providing a list of scenarios, such as those in Table 5, and asking students to either rank them or pick the ones they

care about most. However, how do we operationalize "care"? Is it the scenarios one empathizes with most? Is it the most urgent scenarios to be triaged to the front? Is it the scenarios the students want to learn more about? Is it scenarios they would be willing to put time into, maybe risk their job for?

Additionally, my previous work exposes a notable tension in this context – personally relatable scenarios may be more effective for learning about AI mechanisms and harm [397]; however, if we only teach on personally relevant examples, some vulnerable populations experiencing harm will not be accounted for in these cases. As suggested by in-group favoritism, we are likely to only care about problems immediately affecting us. Given the lack of representation in AI and STEM, this could explain why so many algorithms disproportionately affect marginalized populations. An obvious solution is to increase representation in AI professions – a current goal reflected by many initiatives, including the 21st Century STEM for Girls and Underrepresented Minorities Act, a bill sponsored by Vice President Kamala Harris in 2019. A concurrent solution is to appeal to shared humanity, empathy, and care – demonstrating to learners how AI bias and algorithmic harm affects us all. In other words, can we promote relatability in order to make more people, of all identities, *care*?

As Raji, Scheuerman, and Amironesei [395] writes:

> "We could imagine an **AI ethics class**, for example, in which the narration or participation of **experiential experts** is heavily featured, and students can share their **own reflections on lived experiences with unjust algorithms**, thus connecting personal testimonies to broader studied accounts of the **social and ethical implications** of deployed or speculative systems. The account of practitioners could be included in the class to explain the **logics informing how technical artifacts are designed**, and the kind of affordances or limitations present. Such exercises **build empathy** towards the variety of perspectives present on these issues."

These are normative statements, some of which can be empirically tested. We can evaluate the impacts of including lived experiences of unjust algorithms, and identify where we may need to foster additional empathy and context. In this study, we gave undergraduate Data Science students (N = 172) the option to learn about thirty different algorithmic harm scenarios. For each scenario, we analyzed measures of care via both quantitative and qualitative responses. We operationalize "care" as: a) how personally relatable a scenario is; b) how urgent the scenario seems; c) whether a student sees themselves as good fit to fix the problem; and d) whether or not they would delay deployment on flawed AI software. We then explore the relationships between personal relatability and these other measures of care, identifying both **in-group favoritism** and **in-group responsibility**, as well as qualitatively analysing how students made their decisions about deployment. We ask:

RQ1: How does a learner's personal relatability to algorithmic harm scenarios impact perceived urgency, self-reported belonging, and decision-making?

We find that many of the algorithmic harm scenarios *do* exhibit a positive correlation between relatability and care; though some scenarios cross a sort of moral threshold and are deemed as urgent regardless of relatability. These scenarios are seen as "life or death" urgent, and therefore must immediately be repaired or cancelled. However, the majority of Data Science ethical decisions *are not* life or death. They are often small, seemingly inconsequential, or best-guess tradeoffs that result in harms that may not become apparent until months or years down the line. It is our position that it is a disservice to only focus on life-or-death scenarios of algorithmic harm; we can instead build a curriculum that takes an expansive and empathetic view of a wider range of what constitutes ethical considerations in an AI classroom. We then ask:

RQ2: What are some strategies for effectively integrating AI ethics into AI curricula?

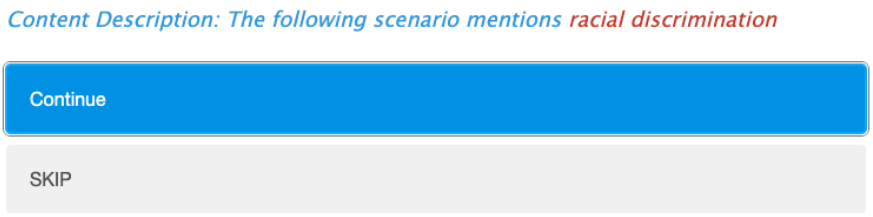
We provide strategies for embedding algorithmic harm scenarios into an AI curriculum based on our empirical findings and the best practice recommendations I developed in Chapter 3. We offer two sample lessons that use our results to inform a buildable series of examples that appeal to both the wide classroom audience, and those most impacted by harm. We match the algorithmic harm scenarios with their technical topics, for example one might discuss the Amazon Hiring Gender Bias scenario when introducing supervised learning, classification, or natural language processing topics. We include empathy-building techniques and ways to incorporate student choice, all embedded within a technical AI/ML curriculum. This ensures that ethics is not othered or left to the end, but integrated alongside technical material as a focus *per lesson*, using real world news stories to speak to student's lived experiences and interests.

Methods

4.2.1 Survey Design

A survey was designed to collect student responses to various algorithmic harm scenarios. In the first study iteration, students had the opportunity to explore all 30 algorithmic harm scenarios listed in 4.2.1.3. Each scenario had the option to be skipped (either for time constraints or sensitive content). Students saw a content description followed by an algorithmic harm scenario, presented in randomized order. Figure 4.2 shows the content description and the scenario for the *Proctoring Software* example. We elaborate more on the sensitive content of this survey in Section 4.2.1.2. For each scenario, students responded to the following questions:

1. Have you heard of this before? (*Yes, No, Maybe*)
2. How relatable is this issue to your personal experiences, identities, or interests? (*Likert Scale*)
3. How urgent is this issue to fix? (*Slider 1-10*)
4. I, more than most others, would be a good fit to address this problem and fix it. (*Likert Scale*)
5. Assume you are part of the team designing and launching the discussed AI tool. **The product sprint ends today, but this issue was just discovered**. The technology is set to be deployed tomorrow. If the deployment is delayed it will inevitably cause profit loss and unhappy executives. **Assuming you have the power to do so, would you:**
    - Stop Deployment to fix the issue
    - Deploy, but work towards a fix
    - Deploy as is, accepting some error
    - Cancel the project and discontinue work



(a) Content warning shown to the student participant with the option to SKIP case

During the pandemic, some universities began relying on automatic proctoring software that tracked and monitored a student's face for cheating on remote exams. However, facial recognition software is known to have low accuracy on dark skinned female faces. Consider a student who is either unable to be detected at all or falsely accused of cheating because the software has this low accuracy.

(b) An algorithmic harm scenario about online proctoring software

Figure 4.2: One of the 30 (a) content descriptions and (b) algorithmic harm scenarios that students saw in the survey.

6. Please provide a short explanation of your reasoning for the above decision you made about deployment. (*Short write-in*)

7. What ideas would you like to share about what could be done to alleviate this specific AI issue? (*Short write-in*)

The survey, created in Qualtrics, consists of five parts: *Introduction, Compensation, and Consent*, *Demographics*, *Pre-Test*, *30 Algorithmic Harm Scenarios*, and a *Post-Test*. The survey consisted of Likert responses, multiple choice, and free-response questions. Students received both extra credit and were entered into a raffle for ten $\$50$ gift cards. Participation was optional; students could choose not to participate did an alternative activity for the same extra credit. The survey was administered in an Introductory undergraduate Data Science course at an R1 institution. The survey lasted approximately 30 minutes, and was immediately followed by a lecture discussing topics of algorithmic harm and algorithmic bias. Data was later analyzed using R, described in Section 4.2.3.

4.2.1.1 Demographics

For each research question, we were interested in how demographics affected student responses. As such, our demographics questionnaire was particularly expansive – including measures about other identities in addition to age, gender, and race. We collected information regarding the following identities: veterans, disability status, LGBTQ+, international students, first-generation Americans, first-generation college students, caregiving responsibilities (i.e. caring for children or elderly adults), and a free-response for other identities the student opted to share. For each of

the demographics questions, there was a *prefer to self-describe* free-text option.

4.2.1.2 Sensitive Content

This study deliberately introduces students to cases of algorithmic harms, and therefore comprises many sensitive topics, requiring additional effort and care in the design and approach. We opted to *not include* any scenarios of algorithmic harm that mentioned physical violence, sexual violence, or white supremacy. While there are certainly examples of algorithmic harm and these topics, it is our position that they should be explored in a setting with more resources available for the students, such as a small directed research group or reading group for opt-in students. The thirty scenarios included in this survey covered topics of racial discrimination, gender discrimination, LGBTQ+ issues, medical diagnoses, conspiracy theories, inequitable access, poverty, mental health, and law enforcement. Each scenario was preceded by a content descriptor and the option to skip that scenario with no penalty (as stated in the consent and instructions), shown in Figure 4.2. Only upon their choosing to Continue was the Scenario displayed to the student participant. Upon SKIP, the participant was brought to the next Content Description. This functionality was included out of respect for student autonomy and agency in sensitive contexts. This study was reviewed by the Human Subjects Division of our university who determined that the proposed activity is human subjects research that qualifies for exempt status – meaning it was determined to be exempt from the federal human subjects regulations, including the requirement for Institutional Review Board (IRB) approval and continuing review.

4.2.1.3 Algorithmic Harm Scenarios

We compiled 30 case studies of algorithmic harm, attempting to represent a wide range of cases, technologies, and types of harm. Each case study is discussed in a news story or empirical research paper, compiled by the author who searched for cases involving various industries, populations, and technologies. In searching for algorithmic harm scenarios, we did consider the identities of the specific student population at our institution – looking for cases that would appeal to international students particularly from Asia and India, financial aid-dependent students, students who experienced online instruction during COVID, and students who identify as LGBTQ+. We also took care to include gender specific scenarios, as to investigate the claim that typical AI instruction may not be adequately connecting with the female student population. This was true for race as well, though Black students are still drastically underrepresented at our institution, resulting in less data for that demographic *(See Section)*. We include some 'popular' or 'high-profile' cases, such as the COMPAS recidivism risk assessment algorithm [21], which we see across current curricula [142, 346]. We also include less popular or more recent cases which we have rarely seen represented and had to be found primarily in news stories, such as when the National Eating Disorders Association (NEDA) used a ChatGPT-powered customer service bot that offered potentially harmful recommendations about calorie counting [499]. For any cases that are not explicitly verified, or for cases that are broader in nature, we used postulating language such as 'may' or 'might' as to not make too strong of claims. For example, radicalization rabbit holes via algorithm are currently contested; with conflicting evidence on how much the algorithms contribute to amplifying misinformation or filter bubble effects [241]. Therefore, we ensure the uncertainty is apparent for that example. Some cases are from several years ago and may have been remedied, but we found them important to include for student critical thinking and relatability. For context, the *Machine Bias* Propublica article about the COMPAS algorithm came out in 2016, and still used in curricula in 2024 to discuss algorithmic bias. Table 14 is a list of all 30 algorithmic harm scenarios, along with their brief description. Students saw a more detailed description in the survey, with the average length across all scenarios being 90 words. Student participants saw the scenarios in a randomized order.

N	Scenario	Description
1	*Proctoring Software*	Facial recognition used while proctoring online exams can falsely accuse students of cheating [524]
2	*Immigration Facial Recognition*	Asian man is told to "open his eyes" by automatic passport photo software in 2016 [1]
3	*Theft Surveillance*	Facial recognition used for theft detection may misidentify suspects leading to wrongful arrest [231, 381, 48, 254]
4	*Matching Mugshots*	Facial recognition software disproportionately and incorrectly matched POC members of Congress to mugshots [448]
5	*Amazon Hiring*	Amazon scraps a hiring tool that disproportionately rejected female candidates [246]
6	*International Hiring*	LinkedIn automatically rejected "out of country" applicants [261]
7	*Financial Aid*	Universities rely on enrollment algorithms which may push students past financial capacity [81]
8	*Loan Approval*	Loan approval algorithms showed racial discrimination against Black applciants [220, 74, 213]
9	*COMPAS*	Recidivism risk assessment algorithm deemed Black offenders as higher risk than white counterparts [21]
10	*Child Welfare*	Allegheny Family Screening Tool is a child welfare algorithm that showed bias against parents with mental illness or disability [181]
11	*Transgender TSA*	Transgender people are stopped disproportionately during TSA screens because their bodies may flag as 'anomolous' by the body scanner [368, 308]
12	*Content Moderation: Activists*	Content moderation algorithms on social media may disproportionately ban transgender, Black, and antiracism educators [212, 305, 401]
13	*Content Moderation: Female Body*	Content moderation algorithms on social media may disproportionately flag female bodies as sexually suggestive even when they follow the rules, and have rules prohibiting female body hair [24, 165]
14	*Ukraine Deepfakes*	Deepfake audio of President Joe Biden saying there will be a draft for American citizens to the war in Ukraine [486, 96]
15	*Content Moderation: PTSD*	Kenyan workers sue Meta for low paid, high volume, and traumatizing content moderation work [337]
16	*Pro-Anorexia Recommender Systems*	Social media recommended content may lead into pro-anorexia content for young teen girls, as brought before the Senate in 2021 [414, 467, 95, 460, 258]
17	*Alcohol and Recommender Systems*	Alcohol ads are marketed to teens and recovering alcoholics via social media recommender algorithms [511, 350, 37]
18	*Radicalization Rabbit Holes*	Youtube recommendations may lead to ideological bubbles or 'echo chambers' towards more fringe or conspiracy theory content [361, 241, 522]

19	*NEDA Chatbot*	National Eating Disorders Association uses a Chat-GPT powered chatbot that recommends cutting calories to someone with an eating disorder [499]
20	*Generative AI Image Bias*	AI images reproduce gender and racial stereotypes for occupations like 'housekeeper', 'CEO', 'doctor', 'janitor', 'teacher',etc [351]
21	*Personalized Disinformation*	Generative AI text can create personalized disinformation that appeals to ones specific tastes and vulnerabilities, a new concern for misinformation researchers [52]
22	*Environmental Impact*	Water and electricity used for data storage, algorithm training, and generative AI have an environmental impact [292, 514]
23	*Health Insurance Bias*	A widely used healthcare algorithm underpredicted how sick Black patients were compared to white counterparts by using historical healthcare costs as a proxy for health [362]
24	*Skin Cancer*	Skin cancer detection software had lower accuracy on darker skin tones [167, 69, 204]
25	*Crisis Hotline*	Crisis hotline sorting algorithm assessing suicide risk had lower accuracy for non-white callers/texters [88, 106]
26	*IBM Oncology*	An IBM cancer research project using Watson recommended fatal drug interactions [163]
27	*Emergency Room Triage*	ER Triage algorithm underpredicted risk for Black patients on a variety of conditions including heart and kidney disease [4, 7, 62]
28	*Voice Assistant Bias*	Voice assistants like Siri or Alexa struggle to respond to slang, accents, and particularly AAVE [293, 276]
29	*Google Translate Bias*	Google translate showed bias for gendered pronouns aligning with stereotypes for occupations like 'nurse', 'doctor', 'engineer', etc. [261, 281]
30	*ChatGPT Privacy Risk*	The amount of data that people are giving to ChatGPT poses a privacy risk, with companies fearing that proprietary code will end up in ChatGPT's training set [47, 205]

Table 14: 30 algorithmic harm scenarios and their brief descriptions, presented in random order in the survey. Students did not see the titles of the scenarios, only the scenarios themselves preceded by a content descriptor and option to SKIP. The wording that students saw was longer, on average 90 words.

4.2.1.4 Pre/Post Test

The survey included a pre- and post- test that measured baseline attitudes towards AI before and after seeing the scenarios of algorithmic harm. Both the pre- and post-test included questions about perceived *benefits* of AI technology as well as potential *harms* from AI technology, with the post-test asking if the participant felt differently after learning about more algorithmic harms. These sections also asked students to define AI that is *'ethical', 'fair', 'responsible'*, and *'transparent'*. The final question asked whether a particular scenario 'stood out' as the most interesting to the participant.

Racial Demographics of Pilot Sample	
Self-Reported Race	*Percentage*
Asian	69%
White	12%
Black	4%
Hispanic	2%
Other	
Self-Describe	13%
Prefer Not to Respond	

Gender Demographics of Pilot Sample	
Self-Reported Gender	*Percentage*
Female	48%
Male	47%
Nonbinary	2%
Prefer Not to Respond	3%

Table 15: Demographic characteristics of pilot sample

4.2.2 Participants

Participants were undergraduate students in an introductory Data Science course at an R1 institution. The total number of participants surveyed was N = 172, however, due to skippable scenarios and time constraints, the mean number of responses per Scenario was N = 52. The mean age was 20 years old. While we recruited respondents of all races, our sample is predominately Asian students, reflecting the demographics in this degree program and the University. We discuss this further in Section , with a plan for quota-based sampling in future work. Student participants reported the following social and demographic characteristics, with race and gender reported in Table 15.

Beyond race and gender we collected other demographic characteristics. **LGBTQ+:** *13% identified as LGBTQ+, and 7% preferred not to respond or marked 'Other'.* **Disability Status:** *80% of respondents were not disabled, 7% were neurodiverse or had a mental disability, 3% had a physical disability, and 10% preferred not to respond.* **First Generation American and International Status:** *40% of respondents were first-generation American, 54% were not. 23% of these students were international students, primarily from China and India.* **First Generation College Students:** *27% were first generation college students, primarily Asian or Black students.* **Primary Caregiving Responsibilities:** *12% had primary caregiving responsibilities, all of whom were either Asian or Black students and primarily (63%) Female.*

4.2.3 Analysis

RQ1: How does a learners personal relatability to algorithmic harm scenarios relate to perceived urgency, self-reported belonging, and decision-making?

In order to answer the research questions proposed in this paper, we use the following analyses. All data was analyzed using R.

4.2.3.1 Relatability and Measures of Care

We are interested in how personal relatability interacts with various measures of care in algorithmic harm scenarios. In order to measure care, we look at perceived urgency on a slider scale, a self-reported 'goodness of fit' as someone who ought to work on the particular problem, and a decision-making proxy regarding deployment of a flawed technology. We evaluate evidence for the following hypotheses:

Hypothesis 1a (H_1a): *When students relate more to a scenario, they find it more urgent.*

We begin by looking at the correlations between relatability and urgency per scenario. Additional tests include evaluating the impact of demographics features on urgency, as well as reporting differing results between scenarios. *If the correlations are positive, there is a positive relationship between relatability and urgency, which we can statistically verify. The strength of the positive relationship can be reported using Spearman's correlation for ordinal data.*

Hypothesis 1b (H_1b): *The more relatable a scenario is, the more the student will see themselves as good fit to fix the problem.*

Our exploratory analysis is similar to the one above, analyzing the response to the question: "I, **more than most others**, would be a good fit to address this problem and fix it" which consisted of an Agreement Likert Scale. We want to know if students see themselves as better suited to solve problems they personally relate to.

Hypothesis 1c (H_1c): *If a scenario is more relatable and/or urgent, a student will be more likely to delay deployment in the case of a problem.*

We investigate if relatability impacts our decision-making proxy, in which students selected how they would handle the algorithmic harm scenario in terms of software deployment. We want to know if students who relate more to a scenario are more likely to find it urgent enough to stop deployment or cancel the project altogether. As discussed, this is a low-stakes proxy for decision-making and will serve as an initial signal of what might happen under real-world constraints.

Looking further into what makes a Scenario *relatable*, and to begin to evaluate in-group favoritism, we look at the mean relatability score per scenario, as well as broken down by demographics characteristics, and whether or not there is a statistically significant difference between relatability scores by either Gender or Race per Scenario. We use a Kruskal-Wallis test for each Scenario to assess differences between groups.

We consider the following hypothesis:

Hypothesis 1d (H_1d): *Students will find cases relatable when they share demographic features with those experiencing harm in the scenario.*

Importantly, we recognize that students may relate to scenarios on deep personal levels in ways that are completely separate from their demographic characteristics. For example, a student with a parent who has cancer may relate to the medical scenarios regardless of their personal demographic information. A student whose cousin was incarcerated may relate to COMPAS or Theft Surveillance, regardless of their own race. Relatability is personal, nuanced; it is the most reflective question we could ask in a quantitative manner to avoid relying on demographics and making assumptions. The relatability question does not specify that relatability must be the students' *own* identities, but can be anything related to their experiences or passions. While exploring associations with demographics might results in observed patterns within the data, we prefer to trust students' own relatability scores and recognize the cause of relatability it could be invisible. This is expanded upon in Table 16 where some students explain their deployment decisions invoking descriptions anchored in personal relatability.

4.2.3.2 Thematic Analysis of Qualitative Decision-Making Responses

Participants selected a deployment decision for each Scenario, then briefly provided "*a short explanation of your reasoning for the above decision you made about deployment*" in a small, free-response text-box. While students were moving quickly through the survey, some answers to this open response question were very detailed. In order to analyze the qualitative responses for the decision-making proxy, we relied on basic affinity diagramming as a preliminary analysis method. In our case, affinity diagramming served as a precursor to thematic analysis, and involved initial sorting of qualitative responses into groups via digital post-it note software. Quotes are sorted into similar categories until saturation; when no new categories are appearing from the data. Categories can then be collapsed or modified. The results presented in 16 are the initial category groups from the author, but need to be verified by additional coders, another limitation of this pilot work.

Results

RQ1: How does a learners personal relatability to algorithmic harm scenarios relate to perceived urgency, self-reported belonging, and decision-making?

4.2.4 Relatability and Perceived Urgency

First, we evaluate support for the hypothesis that perceived urgency of an algorithmic harm scenario will be positively impacted by personal relatability to that scenario (H_1a).

As shown in Figure 4.3, for every scenario, Urgency is positively correlated with Relatability. Overall, regardless of Scenario, the Spearman correlation coefficient (rho, denoted rho) between Urgency and Relatability is $\rho = .286, p < .001$, suggesting a weak positive impact of Relatability on Urgency overall when evaluating the pooled data. However, it is more important to look at the breakdown by Scenario. Some Scenarios have 'flat lines' – a more or less constant Urgency regardless of Relatability. The 'flat line' Scenarios are: Amazon Gender Hiring Bias, Content Moderation PTSD, Crisis Hotline, Emergency Room Triage, Skin Cancer, and Theft Surveillance, with all $\rho \leq .10$. It may be the case that these scenarios are unfair in such an egregious way or speak to shared humanity in a way this quantitative data cannot tell us. On the other hand, some Scenarios have steeper positive linear relationships and higher correlations, indicating a stronger association between Relatability and Urgency. 16 scenarios are observed to have $\rho \geq .3$, at the somewhat arbitrary cutoff for a 'moderate' positive relationship. These include: Alcohol and Recommender Systems, Child Welfare, COMPAS, Environmental Impact, Generative AI Image Bias, Google Translate Gender Bias, Health Insurance Bias, International Hiring Bias, Loan Approval, Personalized Disinformation, Pro-Anorexia Recommender Systems, Proctoring, Radicalization Rabbit Holes, Transgender TSA, Ukraine Deepfakes, and Voice Assistant Bias. The 8 other Scenarios have a weak positive correlation, $\rho >= .15 < .3$. With regard to H_1a, we can determine that for some of the Scenarios it is likely that Relatability impacts Urgency, whereas some Scenarios are High Urgency regardless of personal connection to or situated knowledge of the scenario. It may be the case that these scenarios are regarded as the most egregious, blatant, or life-or-death contexts, leading students to mark them as High Urgency. We elaborate more on this interpretation in the Discussion and provide a perspective on curriculum design that utilizes these empirical findings.

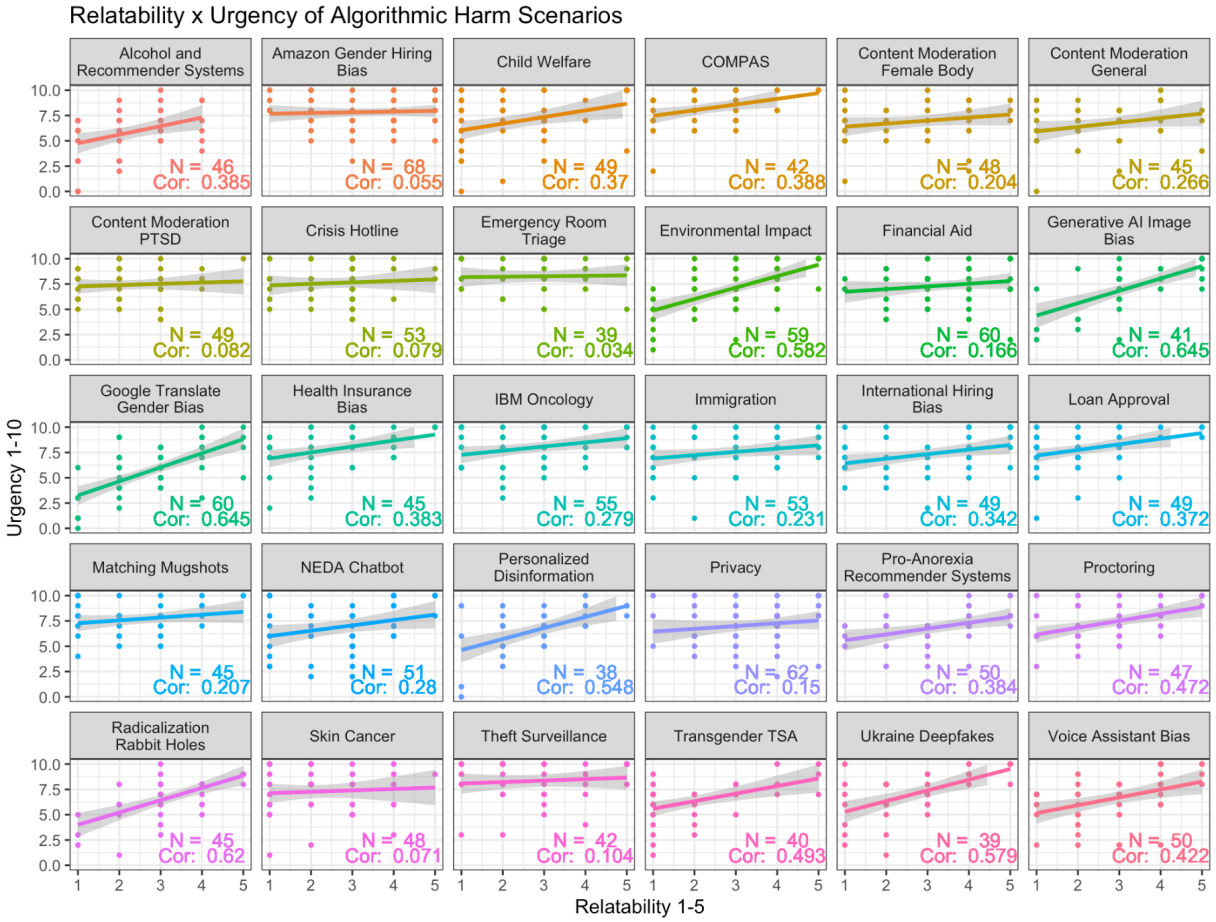


Figure 4.3: Relatability Likert scores (1-5) by Urgency slider scale (1-10) for each of the 30 algorithmic harm scenarios. Because students could SKIP scenarios, number of respondents differs across cases, denoted (N) in the lower, righthand corner, along with the Spearman's correlation coefficient of the relationship. Responses for both variables are ordinal. A steeper slope in the best fit linear model implies a larger effect of Relatability on Urgency, while a flat slope indicates that Urgency remains constant regardless of personal relatability. Note that flat lines tend to be high urgency, indicating a type of 'moral threshold' beyond which relatability does not matter.

4.2.5 Self-Reported Belonging

We asked students to respond to:

> "I, **more than most others**, would be a good fit to address this problem and fix it." *(Agreement 5 point Likert Scale)*

We use data from this prompt to evaluate (H_1b).

Overall, regardless of Scenario, personal Relatability has a moderate positive correlation with Self-Reported Belonging ($rho = 0.535, p < .001$). With regards to our H_1b, we observe that there is likely a positive association between personal relatability and this self-report measure of how "good a fit" one might be to address a problem of algorithmic harm. This suggests that personal relatability plays a role in who will commit to tackling certain problems, see themselves as invested in a problem, and see themselves as worthy of investigating such problems. We could also ask how such commitment is influenced by perceived urgency – in other words, if a problem is highly urgent will people be more likely to commit to working on it? While the two variables do correlate, this may be due to the observed correlation between relatability and belonging. Therefore, we use a multiple linear model and see no significant effect of urgency on belonging (coeff. estimate $= .03, p = .18$), with the relationship being primarily due to relatability (coeff. estimate $= .28, p < .0001$). For each unit of increased relatability on the Likert scale, one's self-reported fit to address the problem increased by .28 units. Some students may find Scenarios to be highly urgent, but not relatable, and therefore may not see themselves as compatible with working on those Scenarios. This provides some evidence in support of *the bystander effect*.

4.2.6 Decision-Making and Ethical Reasoning

Next, we want to see how Relatability and Urgency impact Decision-Making. Recall, we hypothesize about this association in H_1c).

For each Scenario we asked:

> Assume you are part of the team designing and launching the discussed AI tool. **The product sprint ends today, but this issue was just discovered**. The technology is set to be deployed tomorrow. If the deployment is delayed it will inevitably cause profit loss and unhappy executives. **Assuming you have the power to do so, would you:**
>
> - Stop Deployment to fix the issue
> - Deploy, but work towards a fix
> - Deploy as is, accepting some error
> - Cancel the project and discontinue work

As a face validity check, we explore if higher urgency results in increased likelihood of cancelling or stopping the project, as well as the reverse: does lower urgency result in decreased likelihood of deploying "as is"? A linear model reveals a significant negative effect of Urgency for the decision to "Deploy as is, accepting some error" (coeff. estimate $= -2.66p < .0001$); in other words, participants chose to deploy the software without fixing the error when they found the issue not as urgent. This is also reflected in Figure 4.4, which shows higher urgency on average when participants responded to Stop or Cancel the AI product, and a significantly lower urgency for those who chose to deploy "as is".

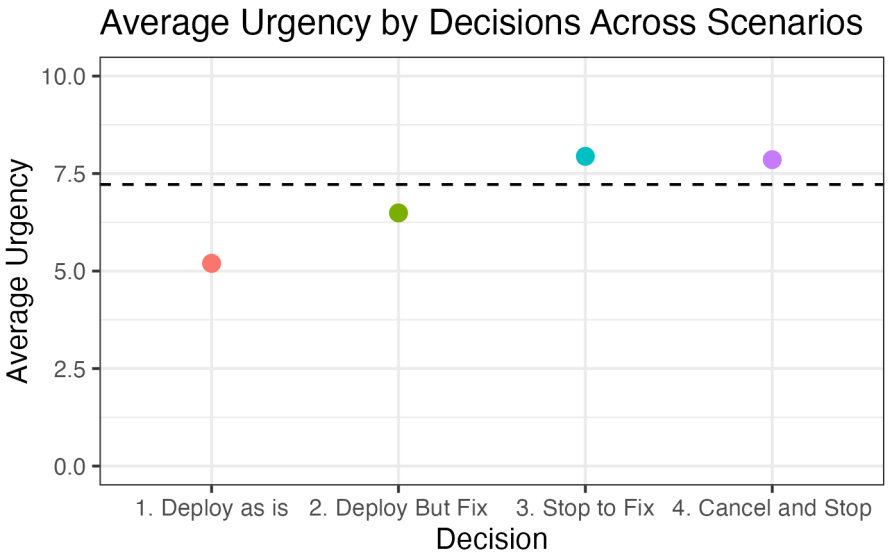


Figure 4.4: Average urgency (across Scenarios) by Decision-making response. The dashed line represents the mean urgency ($\bar{x} = 7.22$) across all Scenarios. We see that for the decision options of "Deploy as is" or "Deploy but fix", mean urgency is below the omnibus mean – this makes sense as when these decisions are made the problem is seen as less pressing or problematic.

We see from a linear model that Relatability is associated with decision-making, with significant positive coefficients of Relatability for *"Cancel the project and discontinue work"* (coeff. estimate $= .47, p < .01$) and *"Stop Deployment to fix the issue"*($coeff.estimate = .30, p < .05$). In other words, when a student relates more to an issue, they likely find it more urgent and are more likely to call for its cancellation or delay its deployment. Keep in mind this measure is a proxy for decision-making and in practice it is much more difficult to stand up to authority and delay a project with multiple stakeholders involved. It may well be that canceling a project, or strongly advocating for such, would be nearly impossible without some risk to one's employment or career trajectory. These results instead give us signals into the severity of these different Scenarios, and how students would *ideally* anticipate reacting in the cases of algorithmic harm.

Looking at the breakdown of decisions directly in Figure 4.5, we see that Proctoring Software and the COMPAS risk assessment algorithm yielded the most "Cancel" decisions. In comparison with something like the Amazon Hiring Scenario, we can infer that students found hiring software to be valuable if the gender bias could be fixed, whereas Proctoring Software and the COMPAS risk assessment seem to be deemed more unnecessary or egregious for what they are beyond just their potential bias. We see this reflected in Scenarios like Emergency Room Triage, where no one said to "Cancel" such an algorithm, and instead encouraged "Stop Deployment to fix the issue". Interpretation of Figure 4.5 may illuminate some ethical tradeoffs that students naturally do – *how useful is the technology vs. how harmful is it?*

This ethical tradeoff reasoning is corroborated by qualitative responses. Table 16 reports key themes and examples from the qualitative written responses to the item: "Please provide a short explanation of your reasoning for the above decision you made about deployment.". We find that participants heavily weigh their decisions in terms of **life or death**, **profit**, **organizational pressures**, **difficulty of repair**, **utilitarian pros and cons**, **frequency of use,** and **personal relatability**. We see some evidence for **care ethics** which we revisit in the Discussion.

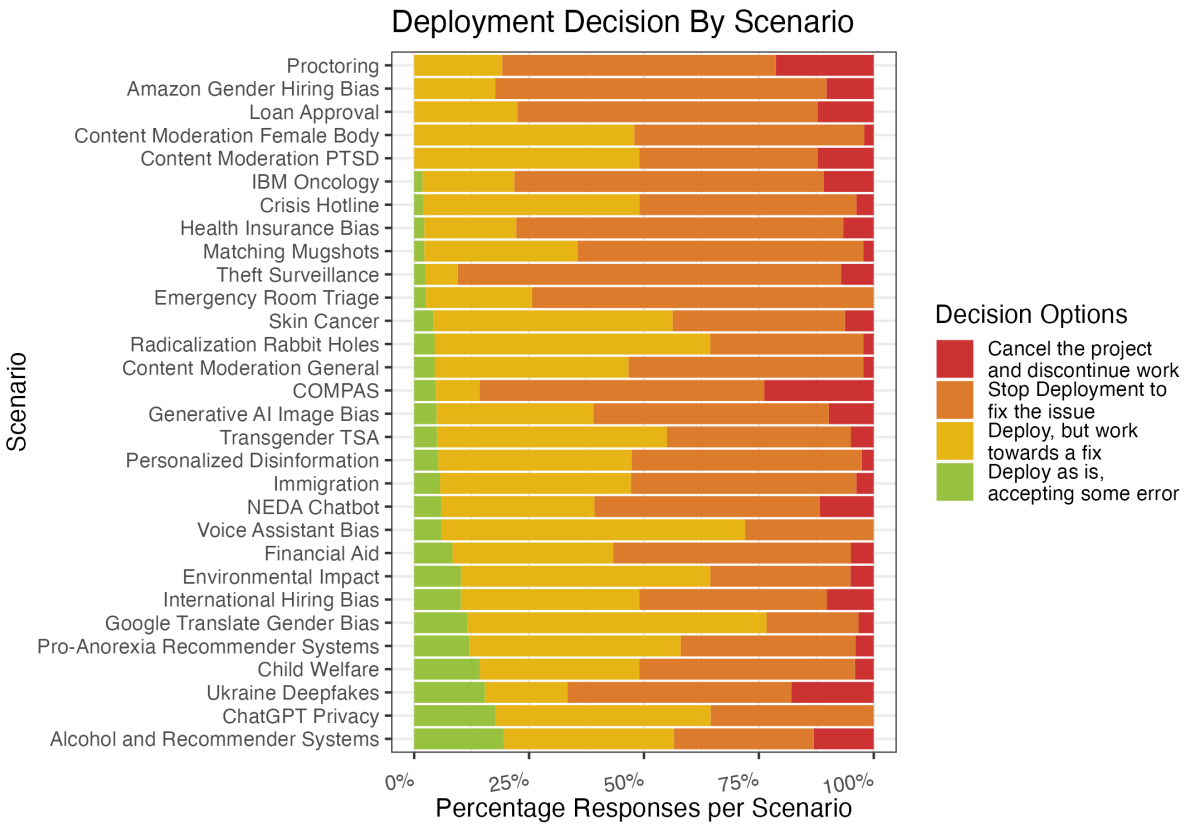


Figure 4.5: Proportion of decision choices made for each Scenario. 5 Scenarios had *zero* "Deploy as is" decisions: Content Moderation PTSD, Content Moderation Female Body, Loan Approval, Amazon Hiring Gender Bias, and Proctoring. 3 Scenarios had *zero* "Cancel the Project" decisions: ChatGPT Privacy, Voice Assistant Bias, and Emergency Room Triage. COMPAS and Proctoring were the Scenarios with the highest proportion of "Cancel" decisions. Alcohol Recommender Systems and ChatGPT Privacy had the highest proportion of "Deploy as is", which may be because students put the onus of responsibility on the user, according to the qualitative data.

Theme	Scenario	Response
Life or Death	*Crisis Hotline*	"Because this is something that should be switched when it comes to **life and death**"
	IBM Oncology	"Recommending potentially fatal drugs that aren't immediately identified by health care providers poses a significant ethical issue. It could lead to preventable **deaths** and trauma."
	Health Insurance	"Concerns **life** = very very very important."
Profit	*Financial Aid*	"If we are not able to cover the **profit loss**, we should deploy it. However, I wouldn't stop deployment to fix the issue."

	Amazon Hiring Gender Bias	"I think we would deploy because I don't want unhappy executives and **profit loss**. Make sure to fix the problem ASAP"
	Emergency Room Triage	"I will still deploy the tool in order to **not risk profit margin**, but at the same time, my team and I will provide the urgency to fix the issue before creating anymore harm to the minorities."
	Loan Approval	"Putting a company's **profits** over people's livelihoods is never something I would want to do"
Organizational Pressures	*Health Insurance*	"I mean it really depends on the **presures** around me to say whether I could actually stop deployment"
	Environmental Impact	"I mean i think the environmental impacts are not worth using the technology. But **no one else cares** so I dont think anyone would stop deployment."
	Loan Approval	"this discrimination is not ok but depends on **work pressure and company leadership** to decide whether the deployment can actually be stopped"
	Radicalization & Recommender Systems	"I think having a plan, and working towards an immediate fix is necessary, so that less people are exposed to the harmful content. However, if I am in such a position, **my line of work and reputation may be at stake**, which prevents me from discontinuing the project in itself"
	Transgender TSA	"Although profit loss and unhappy executives is a large **amount of pressure**, ultimately there is a responsibility to the public to put out a product that does not do harm (to our knowledge) and airport technology is something that people encounter everyday and tons of individuals would be affected by this."

Difficulty of Repair	*Alcohol & Recommender Systems*	"I feel like this specific content about alcohol can be an easier fix so I would stop deployment and fix it before launching. However, the bigger scale of algorithms recommending potentially negative material to people depending on their viewing history is **harder to fix.**"
	Content Moderation PTSD	"I don't want to say cancel the project because that exposes a larger vulnerable population to harmful content, but I'm **not sure how the issue would be fixed**."
	Pro-Anorexia Recommender Systems	"i think social media comes with lots of mental health and self esteem issues, but i also believe that it is an aspect that is **unavoidable**."
Utilitarian Pros & Cons	*Ukraine Deepfakes*	"the **benefits outweight the harms** of the AI even if there are issues"
	Crisis Hotline	"The AI is **still doing good** even if it is not perfectly working as intended" "The service is needed even if its not perfect. So I still think it should exist **even if its not perfect**"
	Environmental Impact	"The benefits of AI is immense. Although there is harm, the **pros are worth the cons** at the moment."
	Voice Assistant Bias	"This device on its own isn't very important. Not everyone needs to have this device. Thus it is okay to **continue with deployment and disregard the issue**."
Frequency of Use	*Voice Assistant Bias*	"If the user relies on these products on the **daily**, it is important to fix them, but they might have more methods in place to resolve such issues already."
	Matching Mugshots	"This is going to be affecting people like on a **daily basis**, so I think it is really terrible they are still using it."

Personal Relatability	*Theft Surveillance*	"im black, this **affects my community**" "The point of AI is to be helpful and not be harmful to people. Since this AI would be harmful to specific groups I would not want it to get out, specifically since **I am have a darker complexion myself**."
	Skin Cancer Detection	"**i am a women of color myself** so this is a problem that affects me and many others"
	Emergency Room Triage	"im a women of color, **this affects my community**"
	Amazon Hiring Gender Bias	"being a women myself in STEM this **affects me** and a ton of others"
	IBM Oncology	"Cancer is a very serious illness. **I, myself** has lost hope and faith in curing cancer. **I have seen to many people around me** who passed .. This tool should have very minimal system errors and a very high accuracy before deploying."
	Voice Assistant Bias	"i think it is a big problem because it is relates to race and **i always meet this kind of problems**"
		"I do not think its as life threatening as other things. **For me** I just often remind myself to use a **whiter sounding voice** when using it."
Care Ethics	*NEDA Chatbot*	"Just one account of this occurance can be **severely damaging** to someone."
	Immigration	"This is racist and embarrassing and causes **emotional distress** to people of color."
	Theft Surveillance	"This sounds **very tragic** and literally putting **someone's life in hell**. need to be fixed before deplyment"

Alcohol & Recommender Systems	"My main concern is the marketing to recovering adults. When marketing substances to kids it can be harmful, but there are also laws in effect that somewhat moderate them being able to access drinking. But for an adult recovering marketing **sets them up to fail**."
Pro-Anorexia Recommender Systems	"Teens, especially females are getting social media at younger and younger ages and this kind of AI can **damage their physical and mental health**."

Table 16: Themes of decision-making in response to different algorithmic harm scenarios and deployment decisions.

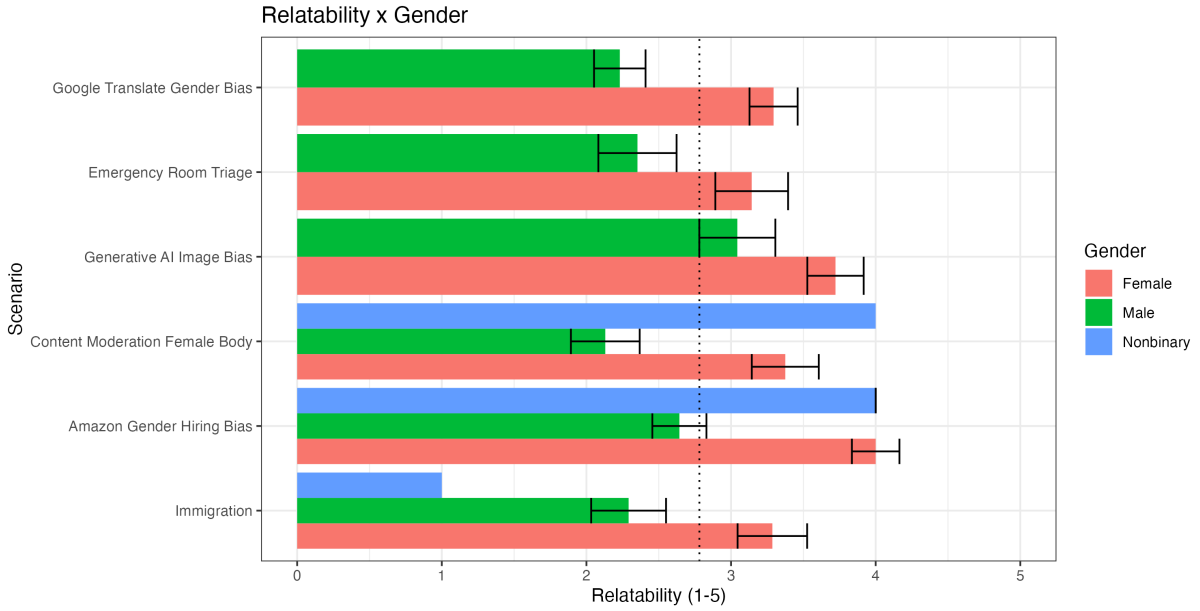
4.2.7 Demographics and Relatability

As part of this investigation we were also able to look at how demographics impacted relatability and urgency to these scenarios. Hagendorff [209] describes the phenomenon of *in-group favoritism*, one of the proposed barriers to putting AI ethics into practice.

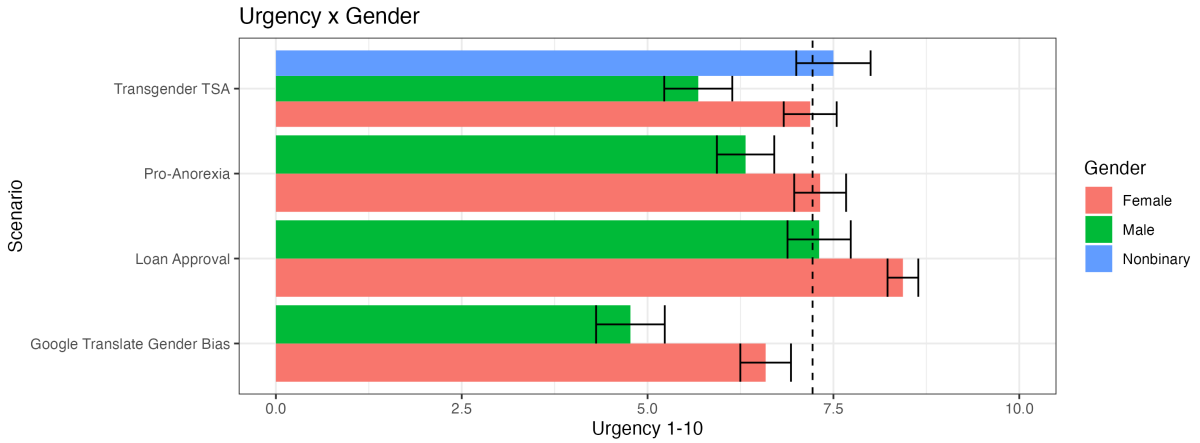
> "*In-group favoritism* causes people to sympathize with others who share their culture, organization, gender, skin color etc. For AI practitioners, this means that AI applications which have negative side-effects on outgroups, for instance the livelihoods of clickworkers in South-east Asia, are rated less ethically problematic than AI applications that would have similar consequences for in-groups. Moreover, the current gender imbalance in the AI field might be prolonged by in-group favoritism in human resource management. In-group favoritism mainly stifles character dispositions like justice and care." [209]

Figure 4.3 already demonstrates some of this, with Relatability positively correlated with Urgency for some Scenarios. We also explicitly test the hypothesis using demographics characteristics (H_1d). Because this sample was not racially balanced, we only evaluate in-group favoritism for Gender. Figure 4.6 shows the Scenarios which had significant effects of Gender on Relatability and Urgency ratings. A Kruskal-Wallis test revealed significant differences (p <.05) between Genders in their Relatability and Urgency ratings. For all the Scenarios, Female ratings were significantly higher than Male. Note that this sample only had N=4 Nonbinary respondents. For Relatability, there were significant differences between Genders for *Google Translate Gender Bias, Emergency Room Triage, Generative AI Image Bias, Content Moderation of Female Body, Amazon Gender Hiring Bias* and *Immigration*. 4/6 of these Scenarios explicitly mention algorithmic harm to women. For Urgency, there was a significant difference between ratings by Gender for the following Scenarios: *Transgender TSA, Pro-Anorexia Recommender Systems, Loan Approval*, and *Google Translate Gender Bias* 3/4 of these Scenarios explicitly mention algorithmic harm to women or gender minorities.

The results shown in 4.6 suggest that demographics do play a role in a student's relatability (and likely perceived urgency) to a scenario, but that they cannot tell the whole story. As stated above, relatability can come in many forms, and it may even be recommended to help students connect to scenarios *beyond* demographics characteristics. We recommend strategies for similarity searching and empathy building in the Discussion, Section 4.3.



(a) Scenarios with significant differences between Relatability by Gender $p < .05$



(b) Scenarios with significant differences between Urgency by Gender $p < .05$

Figure 4.6: Scenarios with significant differences between Gender groups for Relatability (top) and Urgency (bottom). The dashed lines represent overall means for Relatability and Urgency.

4.3 Summary of Findings and Recommendations

RQ1: How does a learners personal relatability to algorithmic harm scenarios relate to perceived urgency, self-reported belonging, and decision-making?

4.3.1 Personal Relatability Impacts Care

Our results indicate that personal Relatability of the learner *does* impact various measures of care when it comes to cases of algorithmic harm. Figure 4.3 shows that for the majority of Scenarios, personal Relatability positively correlates with perceived Urgency, suggesting we care more about things we personally relate to (an interpretation of Hagendorff [209]'s *in-group favoritism*. However, we also see that some Scenarios are perceived as urgent regardless of personal Relatability – supporting the idea of some *moral threshold* for urgency – *some Scenarios are simply too urgent for anyone to ignore*. For example, one of the Scenarios with the least correlation between Relatability and Urgency was the Amazon Hiring Gender Bias scenario, where Amazon had to discontinue a tool used to sift through resumes because it was severely biased against Female candidates. In our study, while Female participants *related* to the scenario more, all genders found it roughly equally urgent to fix, with *no* decisions to "Deploy as is" (See Figure 4.5. We see this pattern of high Urgency regardless of Relatability for scenarios like Theft Surveillance, Content Moderation PTSD, Crisis Hotlines, Emergency Room Triage, and Financial Aid.

On the other hand, the case of Google Translate Gender Bias showed steep correlation between Relatability and Urgency, with a significant effect of Female Gender on both Relatability and Urgency. We see this pattern for other scenarios like Voice Assistant Bias, Transgender TSA, Environmental Impact, Generative AI Image Bias, and Radicalization on YouTube.

Investigation into self-reported belonging ("I, more than most others, would be a good fit to address this problem and fix it") revealed that personal relatability most impacts this score. We did find some evidence supporting *bystander effects*, where the presence of others deters an individual from intervening in a situation of harm. We see this for individuals who rate an issue as high urgency, but low relatability, and therefore a low score for their fit in addressing it. Our stance is that those primarily affected by a harmful technology should take the lead in harm mitigation, but it should not always be left solely to them. Especially due to gender and race disparities in AI, many problems cannot be left to the most impacted alone. Instead, we encourage stakeholder engagement and participatory research; concepts often taught in design courses but less in computer science.

We saw that urgency impacted the decision-making proxy, and that relatability impacted urgency. Therefore, the primary intervention may be empathy-building, as well as practical training on algorithmic auditing throughout a system's lifecycle, as provided by Raji et al. [396]. Judging from our qualitative analysis of reported decision-making reasoning, we find that students were particularly sensitive to "life or death" scenarios, or *clear cut cases of unfairness*, such as the Amazon Hiring tool or Theft Surveillance where someone was wrongly accused of a crime (See Table 16). These *clear cut judgments* lead to high perceived urgency across the board, regardless of personal relatability. However, considering the implications of this result in context motivates the questions: how often will Data Scientists face such grave decisions of life or death? Won't our Data Science students likely end up in jobs with murkier decisions to be made? Likely their decisions will be complex, contextual, with conflicting tradeoffs and stakeholder interests. It is easy to say "do not deploy such an unfair system that causes such blatant harm." It is harder to say "we should use company resources to further investigate this subtle harm that impacts a minority of our user base." And yet, as educators

we have both the freedom and responsibility to urge our students to consider and debate such cases – adequately preparing them to tackle nuanced and insidious algorithmic harms beyond life-or-death scenarios.

One of the themes surfaced for those more nuanced cases of algorithmic harm could be characterized as an *impossibility to fix*: the notion that we cannot stop ChatGPT or deepfakes or the environmental impacts of our technology. However, despite the seemingly impossible, researchers all over the world work on these kinds of issues. Using the fact that this was one of the key barriers for students, we can design our AI courses with this in mind. For example, imagine surveying a class for the problems they find most impossible, and then assigning readings demonstrating efforts from scholars actively working on those issues. We can simultaneously assure our students that they need not act as "ethical unicorns": technologists who do it all, including solve the ethical and social issues on top of the technical implementation [395]. Instead, our goal is to inspire future Data Scientists to grapple with nuanced algorithmic harms; to *see* them with a lens of ethical foresight.

4.3.2 Preparing Students for Practical Decision-Making

Through qualitative analysis (affinity diagramming) we illuminated the ways that students apply ethical reasoning to make decisions about harmful technology. While some students specifically mentioned personal relatability or care ethics, we primarily observed students speaking in terms of utilitarian ethics, life-or-death tradeoffs, or profit analysis. As educators, we have some options for working with those predispositions. On one hand, we might focus our efforts on steering students towards care ethics and empathy as outlined in Section 4.3.6. We may encourage students to grapple with topics that are not life-or-death, but illuminate the very real harms that still occur (reminding students that they will most likely come across more nuanced harms in their careers as opposed to obvious black and white ethical cases dealing with human lives). We can support students to consider conflicting stakeholders, as well as meaningfully engage with what a 'tradeoff' really means. One meaningful strategy for discussing tradeoffs is by analyzing a confusion matrix of false positives and false negatives, asking students to consider the risks and benefits of their classification matrix for each classification task (especially with class imbalance).

On the other hand, we might 'highjack' the predispositions for students to care about life-or-death scenarios and profit margin. We can take scenarios of algorithmic harm and put them in terms of life-or-death, or profit lost. While this is a more dramatic approach, it may be a successful one for some. However, it is important to not only value the dollar amount of potential harms, as this turns ethics into a transactional process rather than a moral one [174].

Through either approach, we can use the themes surfaced in Table 16 to appeal to the pressures that students face when considering ethical decisions. As part of preparing students to consider ethics in their future careers, we must also give them guidance on how to navigate organizational pressures, consider conflicting stakeholders, investigate outliers properly, and assess risk in terms of psychological/physical impact as well as potential dollar amount.

4.3.3 Curriculum Design: Building an AI Curriculum That Successfully Embeds Ethics

RQ2: What are some strategies for effectively integrating AI ethics into AI curricula?

Firstly, we encourage the use of the recommendations provided in Table 8, which build on current literature to provide best practices for AI ethics education. Examples in class should be *specific*, referring to real news stories or AI mistakes. We should aim to be *prescriptive*, informing our students ways to combat and fix the mistakes; so that they are left with some solutions. We encourage an *action-centered* approach, which does not deem our students as unethical people – we instead focus on actions and outcomes that cause harm, not individuals who are "bad" people. We can aim for examples to be *relatable* – in this study we had several examples that appealed directly to students such as the Proctoring example, Financial Aid, or Hiring software. We also recommend some options below for making *any* scenario relatable through empathy-building strategies, with *empathy* being another core characteristic of effective AI ethics education. Examples can be presented *in context*, allowing students to grapple with ideas of societal and organizational pressures. As educators we may not even ask for solutions to the problems, but rather identify the organizational pressures one may face when trying to fix them. Our examples throughout our course ought to be *expansive*, considering many different types of harm and nuanced cases and lesser known cases, many of which are provided in this paper. For student projects and ideas, we can also ask them to *preventatively* look for potential mistakes or harms, teaching them the value of ethical foresight. And finally, we can embed ethics in an *integrated* way, including ethical consideration throughout every lesson instead of othering it as a "soft skill" or a negotiable bookend. This may be done by asking students to include Model cards for their homework assignments, which explicitly asks for a writeup of the model details including ethical considerations. It is our stance that a couple of points per assignment can go towards an ethical consideration discussion question, allowing students to regularly practice the mental check of ethical audits throughout the entire course as opposed to all at the end.

4.3.4 Sample Curriculum

Relying on findings from Figure 4.3, we suggest including both "flat line" and "steep line" scenarios in a lesson that builds upon itself. As discussed above, a "flat line" scenario is one where regardless of personal relatability, students find it highly urgent. It has seemingly crossed a moral threshold and is clearly problematic to the majority of students. This is a starting point to reach everyone in the classroom about a highly urgent issue. For example, the Amazon Hiring Gender Bias was seen as highly urgent across all levels of relatability. This may be due to the clear-cut bias it showed (*discriminating against female candidates*) and the obvious solution (*ensure gender parity in the training data and output*). Following this highly urgent example, we can visit a somewhat related scenario that had a steeper relationship between relatability and urgency. Consider the Google Translate Gender Bias example. It also deals with gender bias, but was seen as urgent among the female participants and less urgent from male participants in our sample. This may be due to the fact that it has less obvious and less material repercussions, and the harm is more psychological and less obvious to those who have not felt gender discrimination in that way. By presenting this scenario in a lesson, we see this as an opportunity to help female students in the classroom feel represented, while also promoting empathy among those who did not originally find it as urgent an issue. Finally, the third component should always be Student Choice. Students can select from scenarios, either the large list or the few we have provided in the Sample Curricula. Students find a news or research article and delve in to a case of algorithmic harm they find personally compelling. They then create a hypothetical Model Card [326] for that technology – detailing its training data, algorithm, and ethical considerations. We demonstrate two options for sample curricula in Tables 17 and 18. We have also provided technical topics that may go along with each of the scenarios, so that educators may fit them in to their already prepared coursework where appropriate.

Lesson Component	Scenario	Technical Keywords	Activity
Urgent to All	*Amazon Hiring Gender Bias*	supervised learning, classification, NLP, logistic regression, Naive Bayes, RNN, class imbalance, training data bias	Technical audit: discussion of training data, class imbalance, and supervised learning
Urgent by Relatability	*Google Translate Gender Bias*	word embeddings, co-occurence, NLP, RNN, Naive Bayes, class imbalance, clustering, training data bias	Empathy-building: causal chain and similarity searching
Student Choice	**Choose From:** *Loan Approval, NEDA Chatbot, Radicalization Rabbitholes, Crisis Hotline, Personalized Disinformation*	supervised learning, NLP, generative language, class imbalance, training data bias	Self-directed research: news story/research article search and hypothetical Model Card [326]

Table 17: An example series of scenarios that could be incorporated into a technical lesson on supervised learning, class imbalance, or Natural Language Processing. We recommend the opening scenario to be one that is deemed urgent by all, accompanied by a technical algorithmic audit. Next, it can be followed by a scenario that steeply correlates with personal relatability, accompanied by an empathy-building exercise. Finally, we value Student Choice to search for a news or research article on the topic and create a hypothetical model card about the technology they chose.

Lesson Component	Scenario	Technical Keywords	Activity
Urgent to All	*Theft Surveillance*	facial recognition, image classification, neural networks, training data bias	Technical audit: discussion of training data, neural networks, and lack of explainability/transparency
Urgent by Relatability	*Generative AI Image Bias*	neural networks, training data bias, text-to-image, image classification	Empathy-building: causal chain and similarity searching
Student Choice	**Choose From:** *Proctoring, Matching Mugshots, Skin Cancer, Transgender TSA, COMPAS*	training data bias, facial recognition, image classification, neural networks	Self-directed research: news story/research article search and hypothetical Model Card [326]

Table 18: An example series of scenarios that could be incorporated into a technical lesson on image classification, neural networks, and/or generative AI images. We recommend the opening scenario to be one that is deemed urgent by all, accompanied by a technical algorithmic audit. Next, it can be followed by a scenario that steeply correlates with personal relatability, accompanied by an empathy-building exercise. Finally, we value Student Choice to search for a news or research article on the topic and create a hypothetical model card about the technology they chose.

4.3.5 Supporting Diverse Students in AI

This paper would be remiss without addressing the representation disparities in AI today. We must keep in mind that the AI field currently underrepresents women and people of color, especially Black and Hispanic individuals [530]. The 2021 AI Index Report reported that 83.9 % of Tenure Track faculty in Computer Science departments at top universities studying AI were Male. For new PhDs with primary focus on AI, they reported 45.6 % White, 22.4% Asian, 2.4% Black, 3.2% Hispanic, 1.6% Multiracial (and 24.8% Unknown). For Tenure Track faculty studying AI, 67% were White and only .6% were Black. With such disparities, it is not a surprise that so many algorithmic harms affect women and racial minorities. It is our opinion that AI educators ought to commit to attracting and retaining diverse students, though the most effective way to do this is through representation and diversity in the teaching teams. So how does this research study aid us in engaging diverse students? Through student voice and student choice, we suggest built-in opportunities for students to explore cases of algorithmic harm that most affect them – signalling to students that their experiences and perspectives matter even in the most technical of settings. By including some algorithmic harm scenarios with steep correlations between relatability and urgency, we allow the impacted students to feel engaged while we promote empathy-building in the others.

4.3.6 Strategies for Empathy Building

It may be the case that in-group favoritism hinders a student from relating to, or investing in, a particular topic that does not seemingly impact them. We certainly cannot care about *everything*, and students should be encouraged to explore their own interests. However, we may be able to use certain strategies to move the needle on personal relatability. We recommend the following empathy-building strategies:

4.3.6.1 Identifying Causal Chains

One effective method for empathy building is to identify a causal chain [67, 520]. Let us take an example from the algorithmic harm scenarios in this paper: the case of PTSD in Kenyan content moderators who sued Meta. Asking students to produce a causal chain of events might work backwards from this lawsuit, and look like Figure 4.7.

Once students have drafted causal chains, they might specifically search for nodes they can relate to, and expand from there.

4.3.6.2 Similarity Searching

The causal chaining exercise is especially useful when paired with a similarity searching exercise. Building from the previous example, while a student may not be a Kenyan crowdsource worker doing content moderation labeling, they might relate specifically to the node of psychological distress from their work, as many students can. They may relate to being underpaid but needing to keep a job, as many international students in Data Science Master's programs must. They can also likely relate to being a social media user who does not want to come across violent imagery. There are several nodes where a student can find similarities. While this may not be possible for every lesson in a jam-packed machine learning course, it may find its place in a final project or group discussion activity.

4.3.6.3 Narrative Storytelling

For the Student Choice elements of the sample curricula, we may encourage students to seek out news articles of algorithmic harms. Many news articles will include narratives, interviews,

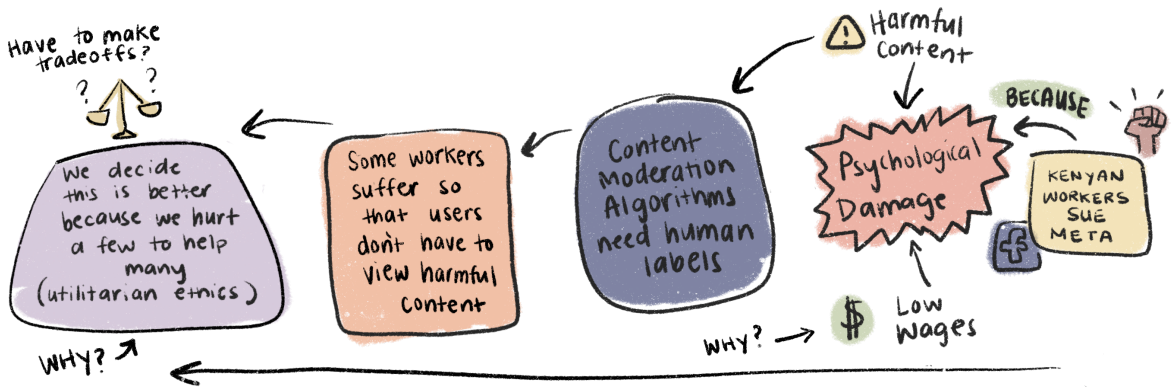


Figure 4.7: An example of reasoning about the causal chain for the case of PTSD in Kenyan content moderators: "Kenyan workers sued Meta, because they experienced psychological trauma from content moderation work. They had to view harmful content in rapid succession, because AI algorithms aren't good enough at detecting it and need to be trained by human labels first. We need people to label harmful content so that others don't have to see it on social media. The company decided it is better to have a small group of people suffer the harms than lots of people, which uses utilitarian ethics."

or specific stories. For example, the viral ProPublica investigation of the COMPAS algorithm opens with the story of an 18 year old girl who is deemed high risk by the algorithmic after stealing (and returning) a bicycle [21]. In the running example from above, we find that news articles covering the Kenyan content moderators also include narrative and quotes from the moderators themselves, giving us insight into their perspective and reasons for the lawsuit. Students can be encouraged to seek out narrative from news articles or qualitative/mixed methods research articles – gaining insight into lived experiences of stakeholders impacted by algorithmic decisions.

4.3.7 Limitations and Cautions

4.3.7.1 Context for Pilot Study

This work as presented is preliminary and consists of results which have not yet been peer-reviewed. This was a pilot study, and the above chapter reports on the *first* iteration of this survey and its results. From this pilot study, we gained insight that allowed us to refine our hypotheses as well as the scope of this work. Further empirical work is needed to verify *in-group favoritism* and its relationship to decision-making. Additionally, decision-making responses could be sorted according to ethical theory, or other measures determined by a larger team. Instead, we simply present the results from an initial affinity diagramming exercise. However, the results of this pilot study are promising, giving us new directions to explore in terms of *in-group favoritism*, student interest in expansive algorithmic harm examples, and barriers they face in their ethical decision-making processes.

4.3.7.2 Survey Design and Burden of Completion

The pilot version of the survey was too long for participants to adequately complete, and this resulted in significant burden of completion for the participants. Many participants chose to SKIP several of the scenarios likely for the sake of time/attention. Students wrote short answers, sometimes repeating their answers simply to finish the survey. It became clear that the survey put too much burden on the participant and needed to be simplified. To address this, we have

redesigned a shorter version of the study, where students select 3 Scenarios to respond to, and then see 3 additional scenarios randomly selected from the remaining 27. Not only does this show us which 3 scenarios are most appealing to students, it also allows us to reach a wider audience for more data collection because the survey is shorter.

4.3.7.3 Demographic Underrepresentation

We were unable to adequately assess the impact of demographics characteristics for this current analysis, due to low response in specific categories. While the demographics component was interesting, we also encourage a broadened view of what Relatability beyond identity can look like. In this specific case, one of the limitations was that this sample was class imbalanced for race, with 69% of the sample identifying as Asian. However, we should explore *why* students related to what they did, beyond demographics characteristics. For example, social media scenarios or the student exam scenario may have appealed to many, beyond racial or gender identity.

4.3.7.4 Generalizability

The results and analysis presented above are preliminary, performed on pilot data which had significant demographic imbalance as well as some missing data due to burden of completion put upon participants. The research questions explored and the pilot findings should be empirically verified with a shorter survey testing *in-group favoritism* as well as purposefully quota-sampled demographics. The findings are promising, as they suggest some cases of algorithmic harm are only perceived as Urgent when students personally relate to the scenario, whereas some cases are seen as morally clear and easier to judge as Urgent. This gives us detailed insight into the baseline ethical reasoning of undergraduate Data Science students. However, these results must be verified and replicated across multiple Data Science cohorts with more balanced demographics and a shorter survey. It may be the case that this method could be used by universities attempting to design their specific AI curricula in their region with their student body – a kind of institution-level personalized curriculum design. If certain cases of algorithmic harm are repeatedly deemed most or least Urgent or Relatable across institutions and regions, we can certainly generalize about those on a larger scale.

4.3.7.5 Lack of Validated Measures

We found the decision-making proxy as well as the potential for empirically validated empathy-building strategies to be the most important for future research, however these are not yet validated measures. The decision-making proxy was a first attempt to measure at how students might behave in a real decision-making scenario. We might improve the validity of such a measure with more forced-choice options or detailed tradeoffs they need to make. This is the same for Urgency and Relatability – in this design, students could rate all 30 scenarios as maximum relatable and maximum urgent. A future version of this study may switch to rankings or voting credits instead. This would mimic a more realistic scenario in which we have limited resources, energy, and motivation to care about *all* of these issues. A forced-choice version may allow us to more accurately pinpoint what is most Urgent and Relatable to our students.

Conclusion

We surveyed undergraduate Data Science students to respond to 30 different real-world news cases of algorithmic harm. These cases ranged from well-known examples such as the COMPAS

recidivism algorithm or the Amazon Hiring software, to lesser known cases of algorithmic harm, for instance social media recommendations drifting towards unwanted dieting content. Interested in how personal relatability impacted various measures of care, we analyzed if students rated scenarios as more urgent when they related to them more. We found that for the majority of scenarios we presented, relatability was positively correlated with urgency. However, some algorithmic harm scenarios were deemed high urgency regardless of personal relatability, seemingly crossing a 'moral threshold' where such mistakes are too egregious to ignore. We saw this reflected in qualitative analysis of why students chose to deploy or delay flawed software: many students mentioned that "life or death" tradeoffs were the most important factor in their decision making. However, we also saw an effect of personal relatability on self-reported belonging to a problem: students who related more saw themselves as a good fit to work on such a problem, whereas those who did not find it relatable suggested others would be a better fit (even if they found it urgent). This suggests we may need scaffolding for supporting more out-group responsibility, while still insisting that those most impacted remain at the center of the design process. We provide two sample lessons that build upon the empirical findings to present algorithmic harm scenarios alongside technical AI instruction. For example, the Amazon Hiring example can be introduced when introducing supervised learning, classification tasks, or natural language processing. It might be followed by looking at gender bias in Google Translate or LLMs, with an opportunity for discussion among students. We also outline how educators can include Student Choice options throughout their curriculum, to allow students to explore their interests in ways that are relevant to both AI technical concepts and their societal impact.

4.4 Chapter Summary and Contributions

✓ Summary

- ✓ I provided recommendations for effective AI ethics education in Chapter 3: AI ethics education ought to be *specific, prescriptivist, action-centered, relatable, empathetic, contextual, expansive, preventative* and *integrated*. This chapter provides empirical results to support some of these claims.
- ✓ One intervention to mitigate algorithmic harm is to more effectively foster empathy in the AI classroom. I introduce care in the context of AI [209], as well as a brief introduction to **Ethics of Care** (or *Care Ethics*)[187, 353]. Ethics of Care in an AI context suggests that centering vulnerable populations and empathy at the forefront is a valid ethical approach to responsible and beneficial AI.
- ✓ Including relatable scenarios of algorithmic harm may foster inclusivity and retain students from underrepresented backgrounds. I explore how personal relatability can go beyond identities of race or gender, but are more expansive and contextual; meaning we can include scenarios of algorithmic harm that relate to students on a personal level for a variety of reasons.
- ✓ In a preliminary study I measure student care towards 30 real-world cases of algorithmic harm.
- ✓ By exploring which cases students relate to, find most urgent, could see themselves working on, and how they make deployment decisions – we gain valuable insight into how to design effective AI curricula that embeds ethics throughout the entire course (rather than leaving ethics to the end [180]).

△ Data Science Tip

Comprehensive Data Science explores a wide variety of systems, algorithms, applications, and types of errors. Algorithmic harm is not limited only to training data bias, but also extends to more expansive views of harm and algorithmic mistakes. By preparing students to look at algorithmic harm from many angles, they will be equipped with a more well-rounded view of the various ways that algorithms impact society. Data Science education ought also to equip future Data Scientists with the communication skills to justify tradeoffs when and where harm occurs – as well as prepare them to make decisions within organizations with conflicting stakeholders. By planning for harm, we can plan for the repair.

★ Contributions

- 30 cases of real-world algorithmic harms that have been evaluated by undergraduate Data Science students in a preliminary study
- Operationalized measures of *care* that can be used in AI ethics contexts: personal relatability, perceived urgency, self-reported goodness of fit, and a decision-making proxy
- findings on the relationship between personal relatability and the urgency of algorithmic harms – revealing that some AI errors cross a moral threshold and are deemed egregious irregardless of personal relatability
- findings that indicate an overreliance on life-or-death stakes in AI ethics decision-making, indicating the need for better support for practical decision-making that Data Scientists deal with on the job
- recommendations and strategies for embedded ethics throughout an AI curriculum in ways that foster both inclusivity and empathy-building (with 2 sample lessons and example activities provided)

Part 5
Resources and Materials for Embedded Ethics in AI Education

"Imagine a world where the Ada Lovelaces of tomorrow grow up to be optimistic and brave about technology and use it to create a new world that is wonderful, whimsical, and a tiny bit weird."

– Linda Liukas, author of *Hello Ruby*

Figure 5.1: Original artwork depicting a Data Science student at their laptop, and the word Educate!

5.1 LEADERS Framework

I had the opportunity to develop the LEADERS Framework for Code.org and their professional learning materials. Inspired by my research as well as my background in education and cognitive science, I was able to put together a simple framework inspired by how I refer to my students: *future Data Science leaders of the world!* These guidelines give a variety of techniques for

discussing difficult topics in the AI classroom. One of the simplest things educators can do to improve their curricula based on the contributions of this dissertation is to choose a scenario presented either in Table 5 or Table 14, and pose the question: *"How did this happen?"*. However, presenting any of these cases of algorithmic harm must be done with sensitivity and care. This chapter presents the LEADERS framework – a set of guidelines for educators who are interested in teaching about cases of algorithmic harm.

As an educator, sometimes with little to no formal training in ethics or social science, it can be daunting to include examples of racism, sexism, exploitation, sexuality, or medical emergency in an AI classroom – and indeed you may decide that you will only present a fraction of the cases I have provided in this dissertation. This framework exists as a jumping off point for those interested in embedding algorithmic harm scenarios into AI curricula. This particular version was written for K-12 instructors, but can be easily applied or extended for more advanced levels.

Verbatim Text

LEADERS Framework as presented in *Our AI Code of Ethics*, *Societal Impacts of Generative AI*, and *Ensuring a Responsible Approach to AI Professional Learning* activities for Code.org.

L - Link to your subject area

AI technology is nearly everywhere! You might be surprised, but AI is not just robots, chatbots, or self-driving cars. It is also: newsfeed organization, Google search, content moderation, financial analytics, audio and video optimization, medical triage, and more! Remember the old phrase "Theres an app for that"? Today its more like *"Theres AI for that"*.

Artificial Intelligence is an umbrella term being applied in almost every sector you can think of, often to amazing results worth celebrating! But there is also the potential for harm in each of these cases. It can be helpful to think of potentials for bias and harm related specifically to your area. You can also discuss AI generally, the workforces involved, the policy regulations, or historical figures.

LEADERS	**Do ✅**	**Don't ❌**
L - link to your subject area	• Speak to your passions and subject area • Find examples of both potentially harmful AI and helpful AI in that space • Encourage students to link concepts to their own experiences	• Assume AI can do everything, even if it is widespread. The errors and biases are widespread too! Often we need people to solve the worlds toughest problems.

E - Errors occur for many reasons

Many cases of unexpected or harmful impacts of AI have errors due to unrepresentative training data. An AI system trained on biased data will produce biased predictions. However, bias can occur from other causes too. For example, a beauty filter makes a face thinner, whiter, and with lighter eyes. This isnt because its missing some data, but because these were the choices made for what defines beauty. What rises to the top of Google search or YouTube may also be because of viral misinformation the system is picking up on engagement metrics or virality rather than the truth of the content. A system can also be gamed one famous example is someone bringing a wagon of cellphones to an intersection, and Google Maps declared a large traffic hold up there.

LEADERS	Do ✅	Don't ❌
E - errors occur for many reasons	• Try to pinpoint why an error happened. Was it a technical error? A social assumption? Gaming of the system? A fluke? • Identify who could have been a part of the development of such technology to avoid the error	• Assume every problem can be fixed with more data. Some algorithms reflect our societal biases and assumptions or need to be fixed with policy and regulation. • Don't assume people who made the mistakes are stupid or uncaring either. We are all capable of making these mistakes.

A - Assume your datapoint is in the room

Cases of AI bias can include: racial discrimination, medical inequity, legal inequity, housing insecurity, gender discrimination, social media manipulation, misinformation, child welfare, education access, food insecurity, etc. Teach as if you are speaking directly to someone affected by the issue, as it may even be the case! We dont know what students may be dealing with at home and should speak about all of these issues as if someone affected is listening.

LEADERS	Do ✅	Don't ❌
A - assume your datapoint is in the room	• Provide content overviews ahead of time (either the class before or at the start of class) • Speak with compassion and a solutions-oriented approach • Recognize that students struggle with difficulties we may not expect	• make inappropriate jokes to lighten the mood • treat the data like it is objective, detached, or infallible • treat outliers like useless datapoints/useless information • use shock value

D - Data sources can be biased

The problems we try to solve and the data we use to solve them can be narrow-minded. For example, trying to use medical predictions from a hospital in Alaska may not generalize or be usable by a hospital based in South Africa. Or selection bias may be excluding people from a community that your students are interested in. Data is not always ethically sourced, and the right questions arent always asked.

Sometimes its necessary to question whether the problem itself is one that we should be trying to solve at all, such as predicting someones professionalism during video interviews for job applicants. Even though its possible to collect data for this situation, is this really a problem we should be solving?

LEADERS	Do ✅	Don't ❌
D - Data sources can be biased	• consider where the data comes from, how it was collected, and for what original purpose • look for if it was ethically sourced • note the year it was collected and the social context of the time (e.g. Boston Housing dataset was collected in a time of severe gerrymandering and segregation (1978))	• assume you always need "more" data, you might need "different" data. • assume all problems are worth solving in the first place, or that AI is a viable way to solve a particular problem.

E - empathy for multiple stakeholders

When interacting with AI systems, empathy is crucial to consider the needs and perspectives of various stakeholders. This means understanding and addressing the concerns of everyone involved, from the users who will interact with the system, to the developers who build it, and the communities who may be impacted by it.

By employing empathy, we can promote more inclusive and effective AI solutions that respect the values and needs of all parties involved. This approach not only results in better user experiences but also ensures that the technology is equitable and just in its operations.

LEADERS	Do ✅	Don't ❌
E - empathy for multiple stakeholders	• Use personas for a wide variety of stakeholders involved, and what their priorities are • Consider the people that the technology works for and who it doesnt • Think about the outliers and how they might occur • Encourage working together with compromise, perhaps in group activities with multiple stakeholders on a team	• Assume your marginalized students will always be the victims of bias, let them take on the role of CEO, programmer, analyst, etc. • Have stakeholders debate when there are issues of identity to be considered, use collaboration instead

R - real world cases

One fear of introducing issues of AI societal impacts and bias in the classroom is that it may be controversial and uncomfortable. One way to overcome this is to rely on real world cases that have happened. At that point, you are teaching history, not an opinion. While it is also important to consider other cases that could occur (preventative brainstorming), you can start by teaching on real-world cases that made the news.

LEADERS	Do ✅	Don't ❌
R - real world cases	• use news stories and have articles prepared to cite • think about similar scenarios that could happen, based on what has happened • keep in mind it may not be labeled as "AI" it may use the term algorithm, machine learning, data science, software	• use fear tactics (i.e. "AI is going to take over the world!") • smooth over cases that have already been fixed (they could happen again in different ways) • use recent cases that haven't been resolved yet, or are distressing in the news, unless you feel prepared

S- Show solutions!

No one wants to feel like their future is doomed - we thrive off hope. For each case of algorithmic bias, there are some solutions that have already worked, and an opportunity to brainstorm possible solutions for the future. Not all solutions are technical fixes either. Sometimes the solution is to have representation in the room.

LEADERS	Do ✅	Don't ❌
S - show solutions!	• provide representation of marginalized scholars, CEOs, technologists, researchers, writers, etc. • provide links to organizations working to solve issues of bias and algorithmic harm • be honest that mistakes will happen, and it takes bravery and accountability to tackle them	• assume that all solutions are technical fixes, or magic "aha!" algorithms. Solutions are often culture or policy-driven. • imply all the problems have already been fixed and won't be repeated in another similar context.

Verbatim Text

Verbatim ends here.

5.2 Case Study: Societal Impacts of Generative AI Activity

I offer *Societal Impacts of Generative AI Activity*, which I co-created with Code.org, for use in your classroom. In this activity, students are grouped into teams of 4, and each represent stakeholers in a Generative AI context. For example, there is a generative AI medical chatbot company – students roleplay as the doctor, the hospital director, the lawyer, and the patient. Each student receives a character card with their perspective on this new generative AI technology, often with both pros and cons (some of which can be seen in Figure 5.3). Each team is given additional news articles and literature to read in order to further understand the scope of the issue. Students procedurally share the benefits and potential harms from the perspective of their stakeholder, and then work together to come up with appropriate solutions. Solutions may be technical, policy-based, company guidelines, advocacy efforts, or a combination of different approaches to ensure responsible generative AI.

There are 9 different scenarios available, each with 4 stakeholder characters. The scenarios include generative AI for *1) writing news and personalized disinformation 2) a beauty app that alters one's personal appearance 3) creative writing in the classroom 4) medical chatbots 5) screenwriting 6) recreating actor likeness 7) AI art and design 8) deepfakes during an election, and 9) writing college essays.* The scenarios are purposefully written to mimic unsolved real-world issues – drawing heavily on news stories from the past year alone. For example, the screenwriter scenario references the 2023 Writer's Strike. The characters for each scenario all present valid points, some which may contradict another stakeholder. This mimics real-world negotiation and tradeoffs. Instead of encouraging students to *debate*, we asked them to *collaborate* and make informed decisions to try to accommodate multiple stakeholders.

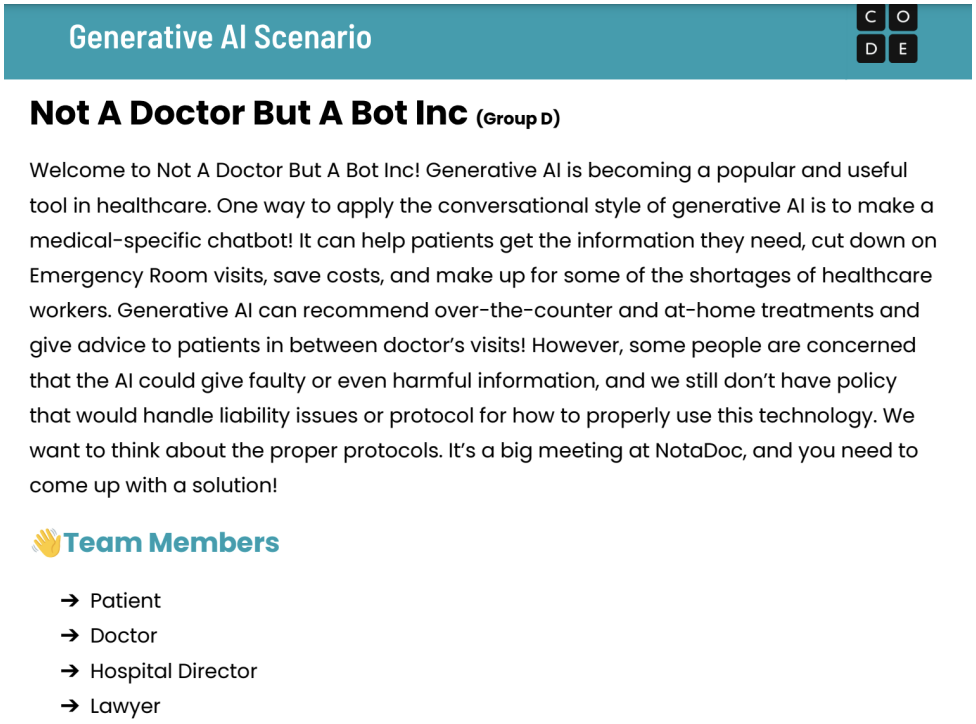


Figure 5.2: One of the Generative AI Scenarios presented to students.

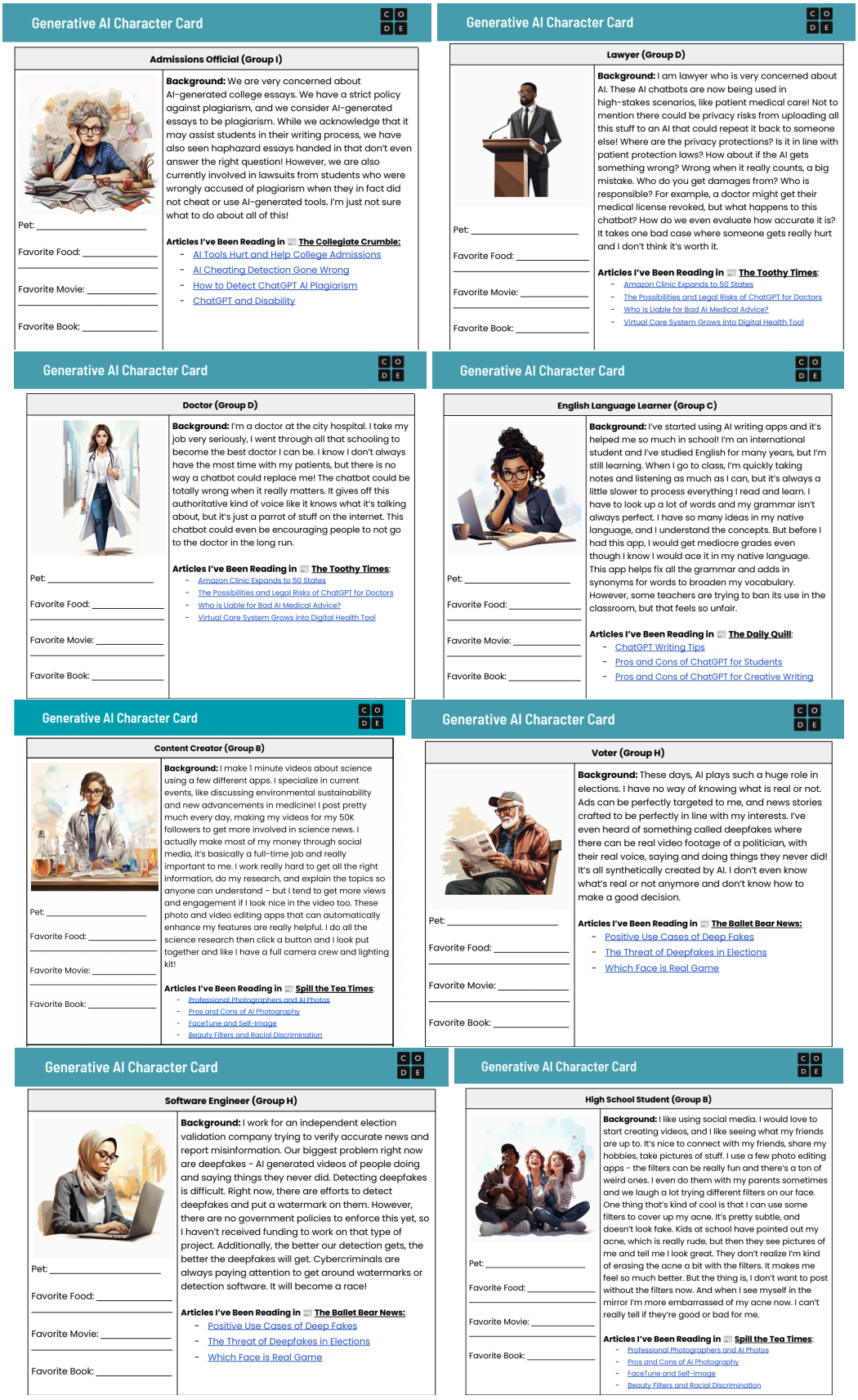


Figure 5.3: A selection of the Character Cards created for the *Societal Impacts of Generative AI* activity. A university admissions official, a lawyer, a doctor, an English language learner, a content creator, a voter, a software engineer, and highschool students – each involved in some kind of generative AI service such as a medical chatbot, a beauty app, or an essay/news writing tool.

5.3 Beyond the Classroom

Figure 5.4: Original artwork depicting an advocate with a megaphone and social media icons, and the word Advocate!

Although this dissertation has focused primarily on applications in the AI classroom, it is important to consider how these findings may be used in other sectors that are increasingly incorporating AI technology as well.

5.3.1 Professional Learning

The techniques presented here can also be applied to professional learning materials. Many companies are now developing AI literacy resources for their employees, to better prepare them to work on AI products as well as integrate AI into their workflows. These professional learning materials typically include the basics of how AI models work, as well as how AI tools can be used on the job. The materials often review the ways that the AI can cause harm or make mistakes, including systematic biases. Not only have I provided many examples of algorithmic mistakes, but I have also demonstrated the benefits of walking through relatable examples of algorithmic experiences in an experiential way – professionals could experiment with AI tools for their own problems and tasks, allowing them to explore how AI works in the context of how it can "break". For generative AI, learners can experiment with prompt engineering or trying to break past AI safety guardrails already in place.

5.3.2 Medicine

AI tools are increasingly being suggested for healthcare professionals – either specifically for diagnosis (such as radiology analysis by machine learning models), or for summarization of doctor's notes by generative AI tools. Another way that AI is impacting medicine is through health misinformation online, which is algortihmically recommended according to your search/watch history and social network, impacting a user's beliefs and trust in medical institutions. We saw this specifically during the COVID-19 pandemic. Many healthcare professionals feel either threatened or overly optimistic about AI in medicine. Therefore, AI literacy should accompany any new AI technology being used by healthcare professionals – detailing the variety of ways that AI can impact medicine and what to do in the cases of misinformation, incorrect results, or biased text summaries or chatbot interactions. This discussion extends to mental health professionals as well, who may be interacting with clients who source information from social media or have developed reliance on AI tools for a variety of tasks. Preparing mental health professionals is something I have done in my outreach work, educating on social media misinformation and algorithmically curated experiences that impact mental health.

5.3.3 Engineering Teams

Education does not end after university. Especially in a field like AI, where technology rapidly shifts and changes, engineers must constantly keep up with new developments. My work can be applied such that engineers not only learn the new technologies but also their potential for harm. Engineers can practice preventative care, data sleuthing, and red-teaming their products and data, using a variety of techniques provided in this dissertation. Exposure to a wide variety of harms may serve as inspiration for features, audits, or data collection that will better serve the team's goals as well as their compliance with budding AI regulation.

5.3.4 Tech Policy and Governance

In order to develop comprehensive policy for AI, we must understand the variety of ways in which it can cause harm. It is crucial that we consider cases of harm *before* they occur, rather than how they may typically coming to our attention through viral news stories or egregious cases. While we cannot prevent every form of harm or every instance, tech policy can proactively engage with the procedures and processes we need to address and repair harm when it occurs. My work not only provides a wide variety of algorithmic harms across industries, but also demonstrates how we can effectively raise awareness about them with stakeholders (both experts and non-experts), translating AI literacy into something personal that resonates across identities. My work is also useful in drafting education policies, generative AI education policies, and a variety of other regulatory efforts.

5.3.5 Libraries and Community Outreach

Education is for all ages and all walks of life – it need not be bounded to a university classroom. Libraries, pop-up events, social media campaigns, art projects, protests, political advocacy, performances, films, games, museums, field trips – so many experiences can be educational. I encourage us all to engage with our fears and hopes for AI technology, bringing it out of the classroom and into public discourse, so that we can all have a say in shaping our AI future.

5.4 Frequently Asked Questions (FAQ)

Figure 5.5: Original artwork depicting a person and a social network, their shirt reads 'I am enough', and the word Liberate!

1. What's missing from AI education right now?

Embedded ethics, of course! The majority of curricula include ethics as a wrap-up topic, which unintentionally 'others' it as something less important, or too broad to engage with in a practical way. We need concrete case studies, auditing techniques, repair strategies, workplace preparation, and lesson plans to integrate ethics throughout the entire AI curriculum, as well as professional development for educators to feel confident teaching sensitive topics. I have provided many of these in this dissertation. As Fiesler, Garrett, and Beard [169] wrote:

> "This reminder – that code is power, and it should be used responsibility– could be part of every computing course, but is arguably most important at the very beginning of the process of learning to code. This strategy might even be a way to combat an 'I'm just an engineer' mindset that ethics is 'someone else's job' by emphasizing its role in computing from day one and then continuing this reminder throughout the curriculum." [169]

2. What are some examples of 'algorithmic harm'?

Check out Table 5 for a list of algorithmic harms across technologies and industries! While certainly not exhaustive, it contains cited examples ranging from wrongful arrests to pro-eating disorder content on social media. You may have heard of many of the algorithmic mistakes: such as the Twitter cropping algorithm that preferred white faces, or how generative AI image tools recreate racial and gender biases when depicting different occupations. I've compiled different kinds of harm, including worker exploitation and environmental impacts of AI. In Table 14, I surveyed undergraduate Data Science students to respond to 30 different algorithmic harm scenarios, including examples of social media recommender systems gone wrong or proctoring software accusing students of cheating.

3. How do I talk about uncomfortable topics in the classroom?

I developed the LEADERS framework to address exactly this! It can be awkward, scary, or potentially controversial to bring up such egregious cases of algorithmic racism, sexism, and harm. However, if there is one thing to keep in mind, *assume your datapoint is in the room*. When we present any kind of dataset, whether we are explicitly discussing harm or not, assume your datapoint is sitting there in the front row. Discussing cancer diagnosis? Consider the student experiencing that with their family. Discussing world health statistics? Our classrooms are beautifully international these days, your datapoints may be right there listening to you talk! Talking about crime statistics, or high BMI, or homelessness, or even job layoffs? Talking about death in any capacity? Be mindful, it will take you far.

4. Many of the harms you listed already got fixed. Why do they matter?

I pose the opposite question: *how would it feel to present cases with no solution?* It may be demoralizing and difficult to learn from in terms of what to do to fix these problems. We learn from history to improve our ethical foresight, and we start to see certain patterns of AI mistakes. For example, many issues of algorithmic bias come from imbalanced or biased training data. We saw this with the Amazon Hiring example or the COMPAS recidivism algorithm. This may serve as a reminder for students to always investigate their data (even if it's just for homework!). I have students write up where the data came from, when it was collected, and other details about the training set. For example, that famously problematic dataset *Boston Housing* is from 1978. What was the historical context of gerrymandering at the time? Finally, we also get the opportunity to learn from the solutions that people came up with. For example, after Google Photos labeled a Black couple as 'gorillas' using automated photo tagging, their solution was to disallow 'gorilla' as a category for search (without completely fixing the underlying algorithmic issue) [207]. They made the decision that such a mistake was so egregious it should be avoided at all costs. Students can analyze the efficacy of such solutions.

5. Some of the harms you listed are *impossible* to fix. What should we do?

I feel this often when people ask me "how else should social media make money other than ad revenue?" or "how do we moderate harmful content online?". It is true that some problems feel too big to tackle, and that by fixing one part of an issue we may create another. That is why my focus as an educator is on teaching students to *justify* their complex decisions from an informed place. I do not believe we adequately prepare our students to grapple with tradeoffs inherent to modeling real-world phenomena with data models. Data Science is a *science*, and we ought to be teaching our students to do research on a variety of stakeholders and technical solutions – and then presenting their conclusions and recommendations in an informed way. It is not nearly enough to just teach algorithms, we must also prepare our students to audit, mitigate, and repair algorithmic harms. We may not be able to get rid of all issues of algorithmic harm, but we can do much better than we are doing now.

6. Why should engineers be dealing with ethics; doesn't someone else do that? (*i.e. a designer, an AI ethicist*)

This question is in response to the idea that we don't need "an ethical unicorn" – someone who is entirely responsible for both the engineering and the ethical consideration. It is true that no one person should be responsible for *any one thing* when it comes to AI technologies. However, I argue that those with technical understanding of the underpinnings of these systems are uniquely situated to discover potential harms. We need not wait until an egregious error makes headlines – instead we can use mistakes from the past and mitigation strategies taught in a Data Science education that prepare Data Scientists to internally audit their systems before their release in a more informed way.

7. How do I get my students to care about ethics?

Perhaps first, make sure there are specific cases *you* care about. Students will feed off of your energy and follow your lead. However, you want to avoid preaching or persuading. Remember, we are *action-centered* not *agent-centered* in our ethical approach. *Actions* can be unethical, our students are not the bad guys. In fact, situating your students as part of the solution is one of the best ways to approach AI ethics. Present these complex cases of algorithmic harm, and genuinely ask for their opinions on both technical and policy solutions. You may rely on our recommendation of causal chaining for empathy-building as one way to surface where they personally relate to such issues (See Figure 4.7). Finally, acknowledging that not all students will care – but the ones you need to reach will hear you.

8. What is the easiest thing I can include in my curricula? I don't have time to do a lot of work.

For each homework assignment, ask students to turn in a Model Card [326]. Model Cards are increasingly used in education, research, and industry for model reporting. The card looks almost like Nutrition Facts for a system – reporting key features of the model and its performance including ethical considerations and what the model should *not* be used for. Categories reported on the Model Card include: basic model details, architecture, training data, performance, potential biases and limitations, and responsible AI considerations.

Another simple thing to include is a self-driven project at the end of the course, where they choose a dataset of interest and ask a research question, answering it with an algorithm they learned in the course. I call this the *Choose Your Own Machine Learning Adventure*, and it directly asks students to consider the ethical implications and societal impact of their projects.

9. We can't just stop deployment of big software projects in giant companies. Give me real advice to tell my students.

You're probably right, and students identified organizational pressures as a key factor in their ethical decision-making, along with potential profit loss. You have a few options. You can lean in to the organizational pressures, and appeal to the customers lost or the profit lost by making egregious AI mistakes. However, I recommend leaning instead on student's natural curiosity and desire to solve problems. Train them to audit algorithms, to spot mistakes, to perform evaluation tests. Inspire them to work together to find issues of bias together. This will get them in the habit of working with a team that values responsible AI. It is true that they likely will not have the power to stop an entire product or delay deployment at the last minute – this is unrealistic. But they *will* have the power to make everyday decisions about training data, acceptable performance thresholds, and generative AI guardrails. Practice now, so they can perform later.

10. In your opinion, what is the most urgent AI issue today?

To me, the most urgent AI issue is the push for products we don't need. Throwing AI at everything is unnecessary and in many cases is not driven by actual stakeholder needs. It seems to be driven by a technological race, market pressure, and "because we can" rather than solving real problems. The market is flooded with potentially harmful generative AI tools right now, when we haven't even addressed issues lurking in non-generative AI systems. Right now, I am most worried about the economic implications of such a massive influx of AI products – that more and more money will go towards potentially useless tools that make content instead of towards social welfare programs, mental health support, or environmental justice.

11. Can you tell me your vision for the future of AI?

I envision a future where technology supports our most meaningful endeavors: building community, connecting people to resources, caring for our planet, improving economic stability, and ensuring safety, housing, healthcare and wellbeing for all. It may sound incredibly lofty, but with all due respect, this is *my* dissertation. I choose to believe against all odds that with enough of us fighting for technology that does good, we will make meaningful waves. For many problems, technology and AI will not even be the solution. We need to stop using tools the problems never asked for. However, I believe we will find great value in harnessing insights from data that help us create a better world. As for generative AI, I have yet to see a use case that takes my breath away. I'll be here waiting, while I train my students to *always* look for the opportunities to do good.

Part 6
Reflections and Conclusion

"most importantly love
like it's the only thing you know how
at the end of the day all this
means nothing
this page
where you're sitting
your degree
your job
the money
nothing even matters
except love and human connection
who you loved
and how deeply you loved them
how you touched the people around you
and how much you gave them"

– Rupi Kaur, milk and honey

6.1 If You Read Anything, Read This

This dissertation has explored strategies to embed cases of algorithmic harm into AI education; situating ethics shoulder-to-shoulder with technical components of AI and machine learning. By including these case studies, we can: a) resonate with and motivate underrepresented students; b) better prepare Data Scientists to design and audit their own systems; and c) surface issues of harm from various stakeholders in order to think more preventatively and mitigate harms before they occur. We have the freedom in academia to design our courses and deliverables – we get the opportunity to put students at the center of the solution: instead of telling them all the ways that AI can go wrong, we ask them how we can create a better world. Through critiquing and repairing algorithms throughout their coursework, students can gain both technical and ethical knowledge simultaneously, which can later be practically applied on the job. Not only do we train good Data Scientists, we train the next generation of AI leaders to code with compassion.

In Chapter 1, I examine the current landscape of Data Science curricula, and find that ethics is often "othered" or seen as a peripheral soft skill, despite numerous calls for integrated ethics and human-centered approaches to AI. I explored the possibility of including student's own personal experiences in how they learn AI technical concepts – testing the hypothesis of how situated learning impacted both technical and advocacy outcomes. I found that students who used their own data to learn about linear regression actually had a boost in both technical comprehension *and* articulated better advocacy arguments in the case of an algorithm gone wrong. This was promising for the idea that using relatable and situated examples could improve learning outcomes – for both mathematical mechanisms and grappling with societal impact of AI systems.

In Chapter 2, I explore the space of algorithmic harms, providing an expansive view of the various ways that AI can make mistakes. Many AI ethics curricula include the COMPAS recidivism algorithm or the sexist Amazon hiring tool; these are great examples of real-world algorithmic bias. However, many cases of algorithmic harm are much more subtle and insidious and less "clear cut" in terms of how to mitigate them. I review complex and nuanced landscapes of algorithmic harms: including my published work on discriminatory content moderation

against vulnerable groups and mental health impacts of recommender systems. I demonstrate that many cases of algorithmic harm are less straightforward, with complicated tradeoffs and conflicting stakeholders.

In Chapter 3, I look at the opportunities for repair when it comes to algorithmic harms. One of the go-to strategies is to follow an AI Code of Ethics or Ethical Guidelines. We see these from large institutions and big tech companies, and these AI principles are often the first line of defense for creating responsible AI. However, research suggests they are ineffective in practice, and difficult to operationalize. I synthesize the literature on the ineffectiveness of AI ethics guidelines, and provide recommendations for more practical AI ethics instruction. Next, I test an intervention for embedding ethics into a technical lesson on recommender systems – using learner's own personal data again to help them surface algorithmic harms on social media. We find that social media users were again able to comprehend both the technical components and the social impact of algorithms when taught in this way; drawing on their lived experiences to surface algorithmic harms and reason about the technicalities of the algorithm simultaneously.

While the above research demonstrated the value of lived experiences for AI ethics instruction, I also had concern about *in-group favoritism*; people tend to only care about what directly impacts them. Especially given the lack of diversity in the AI field, would we only ever care about issues of algorithmic harm that impacted the majority groups? In Chapter 4, I tested the notion of in-group favoritism, as well as explored how students made decisions regarding various cases of algorithmic harms. I found that for some cases of algorithmic harm, there is an apparent 'moral threshold' where all students found the issue to be urgent even if they could not personally relate. These tended to be clear-cut cases of bias that were blatantly unjust. For the more nuanced cases of algorithmic harm, such as Google Translate gender bias, there was a significant effect of gender and relatability on how urgent the issue was perceived. We found that students were much more attuned to issues of life-or-death, but struggled with making decisions about more nuanced cases of harm unless they personally related to them.

In Chapter 5, I provide resources and recommendations to educators interested in embedding ethics throughout their AI courses. I present the LEADERS Framework for teaching AI Ethics, as well as an example activity regarding the Societal Impacts of Generative AI.

Overall, I have demonstrated the benefits of situated learning in AI contexts, with a focus on advocacy and ethical consideration. I have shown that something missing from current AI ethics education is how to grapple with nuanced tradeoffs and justify decisions that may deferentially impact conflicting stakeholders. We will not be able to satisfy all constraints or permanently avoid harms – but we can prepare our future Data Scientists to properly investigate their models with ethical foresight and justify their decisions. We can train our future Data Scientists to repair algorithmic mistakes, and create technology in more human-centered and compassionate ways.

Part 7
References

References

[1] Dec. 2016. URL: https://www.reuters.com/article/us-newzealand-passport-error/new-zealand-passport-robot-tells-applicant-of-asian-descent-to-open-eyes-idUSKBN13W0RL.

[2] Rediet Abebe et al. "Roles for computing in social change". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 252–260.

[3] Crystal Abidin. "Visibility labour: Engaging with Influencers fashion brands and# OOTD advertorial campaigns on Instagram". In: *Media International Australia* 161.1 (2016), pp. 86–100.

[4] Hammaad Adam et al. "Mitigating the impact of biased artificial intelligence in emergency decision-making". In: *Communications Medicine* 2.1 (2022), p. 149.

[5] Adam Mosseri. *Shedding More Light on How Instagram Works*. June 2021. URL: https://about.instagram.com/blog/announcements/shedding-more-light-on-how-instagram-works.

[6] Sray Agarwal and Shashin Mishra. *Responsible AI*. Springer, 2021.

[7] Salman Ahmed et al. "Examining the potential impact of race multiplier utilization in estimated glomerular filtration rate calculation on African-American care outcomes". In: *Journal of general internal medicine* 36 (2021), pp. 464–471.

[8] Sara Ahmed. "Orientations: Toward a queer phenomenology". In: *GLQ: A journal of Lesbian and Gay Studies* 12.4 (2006), pp. 543–574.

[9] *AI vs. Machine Learning vs. Data Science for Industry*. Nov. 2022. URL: https://braincube.com/resource/manufacturing-ai-vs-machine-learning-vs-data-science/.

[10] Kathryn Alesandrini and Linda Larson. "Teachers bridge to constructivism". In: *The clearing house* 75.3 (2002), pp. 118–121.

[11] Dimitris Alimisis and Chronis Kynigos. "Constructionism and robotics in education". In: *Teacher education on robotic-enhanced constructivist pedagogical methods* (2009), pp. 11–26.

[12] Ali Alkhatib. "To live in their utopia: Why algorithmic systems create absurd outcomes". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–9.

[13] Ali Alkhatib and Michael Bernstein. "Street-level algorithms: A theory at the gaps between policy and decisions". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–13.

[14] Allison McDonald et al. ""It's stressful having all these phones": Investigating Sex Workers' Safety Goals, Risks, and Practices Online". In: *Proceedings of the 30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, Aug. 2021. URL: https://www.usenix.org/conference/usenixsecurity21/presentation/mcdonald.

[15] Peter Allmark. “Can there be an ethics of care?” In: *Journal of medical ethics* 21.1 (1995), pp. 19–24.

[16] Roya Jafari Amineh and Hanieh Davatgari Asl. “Review of constructivism and social constructivism”. In: *Journal of social sciences, literature and languages* 1.1 (2015), pp. 9–16.

[17] Paul Anderson et al. “An undergraduate degree in data science: curriculum and a decade of implementation experience”. In: *Proceedings of the 45th ACM technical symposium on Computer science education*. 2014, pp. 145–150.

[18] Ruth E. Anderson et al. “A Data Programming CS1 Course”. In: *Proceedings of the 46th ACM Technical Symposium on Computer Science Education*. SIGCSE ’15. Kansas City, Missouri, USA: ACM, 2015, pp. 150–155. ISBN: 978-1-4503-2966-8. DOI: 10.1145/2676723.2677309. URL: http://doi.acm.org/10.1145/2676723.2677309.

[19] W Carsten Andresen. “Research Note: Comparing the Gay and Trans Panic Defenses”. In: *Women & Criminal Justice* 32.1-2 (2022), pp. 219–241.

[20] Andrew Hutchinson. “Instagram Looks to Crackdown on Bots with New Review and ID Process”. In: *Social Media Today* (Aug. 2020). URL: https://www.socialmediatoday.com/news/instagram-looks-to-crackdown-on-bots-with-new-review-and-id-process/583491/.

[21] Julia Angwin et al. “Machine bias”. In: *Ethics of data and analytics*. Auerbach Publications, 2022, pp. 254–264.

[22] Aaron Appelle. “Will businesses or laws and regulations ever prioritise environmental sustainability for AI systems?” In: (2023).

[23] Carolina Are. “How Instagrams algorithm is censoring women and vulnerable users but helping online abusers”. In: *Feminist media studies* 20.5 (2020), pp. 741–744.

[24] Carolina Are. “The Shadowban Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram”. In: *Feminist Media Studies* (2021), pp. 1–18.

[25] Alejandro Barredo Arrieta et al. “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”. In: *Information Fusion* 58 (2020), pp. 82–115.

[26] Andrew Arsht and Daniel Etcovitch. “The human cost of online content moderation”. In: *Harvard Law Review Online, Harvard University, Cambridge, MA, USA. Retrieved from https://jolt. law. harvard. edu/digest/the-human-cost-ofonline-content-moderation* (2018).

[27] Charles Arthur. *Tech giants may be huge, but nothing matches big data, 2013*. 2021.

[28] Christy Ashley and Tracy Tuten. “Creative strategies in social media marketing: An exploratory study of branded social content and consumer engagement”. In: *Psychology & Marketing* 32.1 (2015), pp. 15–27.

[29] Erica Weintraub Austin and Stacey JT Hust. “Targeting adolescents? The content and frequency of alcoholic and nonalcoholic beverage ads in magazine and video formats November 1999–April 2000”. In: *Journal of Health Communication* 10.8 (2005), pp. 769–785.

[30] Imran Awan. “Islamophobia and Twitter: A typology of online hate against Muslims on social media”. In: *Policy & Internet* 6.2 (2014), pp. 133–150.

[31] Joseph B Bak-Coleman et al. “Combining interventions to reduce the spread of viral misinformation”. In: *Nature Human Behaviour* (2022), pp. 1–9.

[32] Mad Price Ball. *Open Humans*. https://www.openhumans.org/.

[33] Albert Bandura and David C McClelland. *Social learning theory*. Vol. 1. Englewood cliffs Prentice Hall, 1977.

[34] Natã M Barbosa and Monchu Chen. "Rehumanized crowdsourcing: a labeling framework addressing bias and ethics in machine learning". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–12.

[35] Solon Barocas, Moritz Hardt, and Arvind Narayanan. "Fairness in machine learning". In: *Nips tutorial* 1 (2017), p. 2.

[36] Thomas E Barone. "Beyond theory and method: A case of critical storytelling". In: *Theory into practice* 31.2 (1992), pp. 142–146.

[37] Adam E Barry et al. "Alcohol marketing on Twitter and Instagram: Evidence of directly advertising to youth/adolescents". In: *Alcohol and alcoholism* 51.4 (2016), pp. 487–492.

[38] Austin Cory Bart et al. "Computing with corgis: Diverse, real-world datasets for introductory computing". In: *ACM Inroads* 8.2 (2017), pp. 66–72.

[39] Austin Cory Bart et al. "Reconciling the Promise and Pragmatics of Enhancing Computing Pedagogy with Data Science". In: *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*. ACM. 2018, pp. 1029–1034.

[40] Kim Bartholomew. "From childhood to adult relationships: Attachment theory and research." In: (1993).

[41] Kim Bartholomew and Leonard M Horowitz. "Attachment styles among young adults: a test of a four-category model." In: *Journal of personality and social psychology* 61.2 (1991), p. 226.

[42] Jo Bates et al. "Integrating FATE/critical data studies into data science curricula: where are we going and how do we get there?" In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 425–435.

[43] Ben Baumer. "A data science course for undergraduates: Thinking with data". In: *The American Statistician* 69.4 (2015), pp. 334–342.

[44] Tom L Beauchamp and James F Childress. *Principles of biomedical ethics*. Oxford University Press, USA, 2001.

[45] Ghazaleh Beigi and Huan Liu. "Privacy in social media: Identification, mitigation and applications". In: *arXiv preprint arXiv:1808.02191* (2018).

[46] Emily M Bender, Dirk Hovy, and Alexandra Schofield. "Integrating ethics into the NLP curriculum". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. 2020, pp. 6–9.

[47] Emily M Bender et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021, pp. 610–623.

[48] TJ Benedict. "The computer got it wrong: Facial recognition technology and establishing probable cause to arrest". In: *Wash. & Lee L. Rev.* 79 (2022), p. 849.

[49] Seyla Benhabib. *Situating the self: Gender, community, and postmodernism in contemporary ethics*. Psychology Press, 1992.

[50] Ruha Benjamin. "Race after technology: Abolitionist tools for the new jim code". In: *Social Forces* (2019).

[51] Roukaya Benjelloun and Yassine Otheman. “Psychological distress in a social media content moderator: A case report”. In: (2020).

[52] Thor Benson. *This disinformation is just for you*. Aug. 2023. URL: https://www.wired.com/story/generative-ai-custom-disinformation/.

[53] Alessandro Bessi. “Personality traits and echo chambers on facebook”. In: *Computers in Human Behavior* 65 (2016), pp. 319–324.

[54] Rahul Bhargava and Catherine DIgnazio. “Designing tools and activities for data literacy learners”. In.

[55] Ismail Bile Hassan et al. “Data science curriculum design: A case study”. In: *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. 2021, pp. 529–534.

[56] Reuben Binns et al. “Like trainer, like bot? Inheritance of bias in algorithmic content moderation”. In: *International conference on social informatics*. Springer. 2017, pp. 405–415.

[57] Abeba Birhane. “Algorithmic injustice: a relational ethics approach”. In: *Patterns* 2.2 (2021).

[58] Sophie Bishop. “Algorithmic experts: Selling algorithmic lore on YouTube”. In: *Social Media+ Society* 6.1 (2020), p. 2056305119897323.

[59] Sophie Bishop. “Anxiety, panic and self-optimization: Inequalities and the YouTube algorithm”. In: *Convergence* 24.1 (2018), pp. 69–84.

[60] Sophie Bishop. “Managing visibility on YouTube through algorithmic gossip”. In: *New media & society* 21.11-12 (2019), pp. 2589–2606.

[61] Danielle Blunt and Ariel Wolf. “Erased: The impact of FOSTA-SESTA and the removal of Backpage on sex workers”. In: *Anti-trafficking review* 14 (2020), pp. 117–121.

[62] Sean Boley et al. “Investigating racial disparities within an emergency department rapid-triage system”. In: *The American Journal of Emergency Medicine* 60 (2022), pp. 65–72.

[63] Tolga Bolukbasi et al. “Man is to computer programmer as woman is to homemaker? debiasing word embeddings”. In: *Advances in neural information processing systems* 29 (2016).

[64] Ludovico Boratto, Gianni Fenu, and Mirko Marras. “The effect of algorithmic bias on recommender systems for massive open online courses”. In: *European Conference on Information Retrieval*. Springer. 2019, pp. 457–472.

[65] Pepe Borrás Pérez. “Facebook doesnt like sexual health or sexual pleasure: Big techs ambiguous content moderation policies and their impact on the sexual and reproductive health of the youth”. In: *International Journal of Sexual Health* 33.4 (2021), pp. 550–554.

[66] Nick Bostrom. “The Vulnerable World Hypothesis”. In: ().

[67] Tammy Bourg et al. “The effects of an empathy-building strategy on 6th graders’ causal inferencing in narrative text comprehension”. In: *Poetics* 22.1-2 (1993), pp. 117–133.

[68] Roz Boustead and Mal Flack. “Moderated-mediation analysis of problematic social networking use: The role of anxious attachment orientation, fear of missing out and satisfaction with life”. In: *Addictive Behaviors* 119 (2021), p. 106938.

[69] Gabriella Brancaccio et al. “Artificial Intelligence in Skin Cancer Diagnosis: A Reality Check”. In: *Journal of Investigative Dermatology* (2023).

[70] Stacy M Branham et al. “Co-creating & identity-making in CSCW: revisiting ethics in design research”. In: *Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing*. 2014, pp. 305–308.

[71] Dawn B Branley and Judith Covey. “Pro-ana versus pro-recovery: A content analytic comparison of social media users communication about eating disorders on Twitter and Tumblr”. In: *Frontiers in psychology* 8 (2017), p. 1356.

[72] Tim Brennan, William Dieterich, and Beate Ehret. “Evaluating the predictive validity of the COMPAS risk and needs assessment system”. In: *Criminal Justice and Behavior* 36.1 (2009), pp. 21–40.

[73] Inge Bretherton. “Attachment theory: Retrospect and prospect”. In: *Monographs of the society for research in child development* (1985), pp. 3–35.

[74] Liming Brotcke. “Time to assess bias in machine learning models for credit decisions”. In: *Journal of Risk and Financial Management* 15.4 (2022), p. 165.

[75] Anna Brown et al. “Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-making in Child Welfare Services”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM. 2019, p. 41.

[76] John Seely Brown, Christian Heath, and Roy Pea. *Vygotsky’s educational theory in cultural context*. Cambridge University Press, 2003.

[77] Robert J Brunner and Edward J Kim. “Teaching data science”. In: *Procedia Computer Science* 80 (2016), pp. 1947–1956.

[78] Taina Bucher. “The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms”. In: *Information, communication & society* 20.1 (2017), pp. 30–44.

[79] Taina Bucher. “Want to be on the top? Algorithmic power and the threat of invisibility on Facebook”. In: *New media & society* 14.7 (2012), pp. 1164–1180.

[80] Joy Buolamwini and Timnit Gebru. “Gender shades: Intersectional accuracy disparities in commercial gender classification”. In: *Conference on fairness, accountability and transparency*. PMLR. 2018, pp. 77–91.

[81] Lilah Burke. *Why colleges are using algorithms to determine financial aid levels*. Sept. 2023. URL: https://www.highereddive.com/news/colleges-enrollment-algorithms-aid-students/692601/.

[82] Ed Burns. *SearchEnterpriseAI Current state of AI is poorly understood by the public*. https://searchenterpriseai.techtarget.com/opinion/Current-state-of-AI-is-poorly-understood-by-the-public. 2018.

[83] Jenna Burrell. “How the machine thinks: Understanding opacity in machine learning algorithms”. In: *Big Data & Society* 3.1 (2016), p. 2053951715622512.

[84] Jenna Burrell et al. “When Users Control the Algorithms: Values Expressed in Practices on Twitter”. In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019), pp. 1–20.

[85] Carole Cadwalladr. “Google is not justa platform. It frames, shapes and distorts how we see the world”. In: *The Guardian* 11.12 (2016), p. 2016.

[86] Joseph A Calandrino et al. “" You might also like:" Privacy risks of collaborative filtering”. In: *2011 IEEE symposium on security and privacy*. IEEE. 2011, pp. 231–246.

[87] Erik Calleberg. *Making Content Moderation Less Frustrating: How Do Users Experience Explanatory Human and AI Moderation Messages*. 2021.

[88] Yana Calou and Hannah Zeavin. *Whos listening when you call a crisis hotline?* Apr. 2022. URL: https://slate.com/technology/2022/04/crisis-lifelines-surveillance-geolocation-algorithms.html.

[89] Matthew Carrasco and Andruid Kerne. "Queer Visibility: Supporting LGBT+ Selective Visibility on Social Media". en. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Montreal QC Canada: ACM, Apr. 2018, pp. 1–12. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3173824. URL: https://dl.acm.org/doi/10.1145/3173574.3173824 (visited on 08/19/2022).

[90] Samuel Carton, Qiaozhu Mei, and Paul Resnick. "Feature-Based Explanations Don't Help People Detect Misclassifications of Online Toxicity". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14. 2020, pp. 95–106.

[91] Casey Newton. "Here's how Twitter's new algorithmic timeline is going to work". In: *The Verge* (Feb. 2016). URL: https://www.theverge.com/2016/2/6/10927874/twitter-algorithmic-timeline.

[92] M Castelli. *Introduction to critical race theory and counter-storytelling*. 2021.

[93] Eva Cernadas and Manuel Fernández-Delgado. "Embedded ethics to teach machine learning courses: an experience". In: *2021 XI International Conference on Virtual Campus (JICV)*. IEEE. 2021, pp. 1–4.

[94] Abhijnan Chakraborty et al. "Who makes trends? understanding demographic biases in crowdsourced recommendations". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 11. 2017.

[95] Stevie Chancellor et al. "# thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities". In: *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. 2016, pp. 1201–1213.

[96] Reuters Fact Check. *Fact check: Video of Joe Biden calling for a military draft was created with AI | Reuters*. Oct. 2023. URL: https://www.reuters.com/fact-check/video-joe-biden-calling-military-draft-was-created-with-ai-2023-10-19/.

[97] Howard Chen. *MSandE 238 Blog Public perception of artificial intelligence*. https://mse238blog.stanford.edu/2018/07/howachen/public-perception-of-artificial-intelligence/. 2018.

[98] Janet X Chen et al. "Trauma-Informed Computing: Towards Safer Technology Experiences for All". In: *CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–20.

[99] Mingming Cheng and Carmel Foley. "Algorithmic management: The case of Airbnb". en. In: *International Journal of Hospitality Management* 83 (Oct. 2019), pp. 33–36. ISSN: 02784319. DOI: 10.1016/j.ijhm.2019.04.009. URL: https://linkinghub.elsevier.com/retrieve/pii/S0278431919300143 (visited on 12/05/2022).

[100] Marc Cheong, Kobi Leins, and Simon Coghlan. "Computer science communities: Who is speaking, and who is listening to the women? Using an ethics of care to promote diverse voices". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 106–115.

[101] Ann-Dorte Christensen and Sune Qvotrup Jensen. “Doing intersectional analysis: Methodological implications for qualitative research”. In: *NORA-Nordic Journal of Feminist and Gender Research* 20.2 (2012), pp. 109–125.

[102] Matteo Cinelli et al. “The COVID-19 social media infodemic”. In: *Scientific reports* 10.1 (2020), pp. 1–10.

[103] William J Clancey. “A tutorial on situated learning”. In: *Proceedings of the international conference on computers and education (Taiwan)*. Charlottesville, VA: AACE. 1995, pp. 49–70.

[104] William S Cleveland. “Data science: an action plan for expanding the technical areas of the field of statistics”. In: *International statistical review* 69.1 (2001), pp. 21–26.

[105] Kirsti K Cole. “It’s like she’s eager to be verbally abused: Twitter, trolls, and (en) gendering disciplinary rhetoric”. In: *Feminist Media Studies* 15.2 (2015), pp. 356–358.

[106] R Yates Coley et al. “Racial/ethnic disparities in the performance of prediction models for death by suicide after mental health visits”. In: *JAMA psychiatry* 78.7 (2021), pp. 726–734.

[107] Combahee River Collective. “The Combahee river collective statement”. In: *Home girls: A Black feminist anthology* 1 (1983), pp. 264–274.

[108] Patricia Hill Collins. “Comment on Hekman’s" Truth and method: Feminist standpoint theory revisited": Where’s the power?” In: *Signs: Journal of Women in Culture and Society* 22.2 (1997), pp. 375–381.

[109] Kelley Cotter. “Shadowbanning is not a thing: black box gaslighting and the power to independently know and credibly critique algorithms”. In: *Information, Communication & Society* (2021), pp. 1–18.

[110] Kelley Cotter. “Shadowbanning is not a thing: black box gaslighting and the power to independently know and credibly critique algorithms”. In: *Information, Communication & Society* 26.6 (2023), pp. 1226–1243.

[111] Kelley Cotter. “Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram”. In: *New Media & Society* 21.4 (2019), pp. 895–913.

[112] Kelley Marie Cotter. *Critical Algorithmic Literacy: Power, Epistemology, and Platforms*. Michigan State University, 2020.

[113] National Research Council et al. *How people learn: Brain, mind, experience, and school: Expanded edition*. National Academies Press, 2000.

[114] Tom Crick. “Computing education: An overview of research in the field”. In: *London: Royal Society* (2017).

[115] Catherine D’Ignazio and Rahul Bhargava. “DataBasic: Design principles, tools and activities for data literacy learners”. In: *The Journal of Community Informatics* 12.3 (2016).

[116] Catherine D’Ignazio and Lauren F Klein. *Data feminism*. Mit Press, 2020.

[117] Catherine D’Ignazio et al. “Toward equitable participatory design: Data feminism for CSCW amidst multiple pandemics”. In: *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. 2020, pp. 437–445.

[118] Maria Chiara DArienzo, Valentina Boursier, and Mark D Griffiths. “Addiction to social media and attachment styles: a systematic literature review”. In: *International Journal of Mental Health and Addiction* 17.4 (2019), pp. 1094–1118.

[119] Catherine DIgnazio and Rahul Bhargava. “Approaches to building big data literacy”. In.

[120] Catherine DIgnazio and Lauren F Klein. “Feminist data visualization”. In: *Workshop on Visualization for the Digital Humanities (VIS4DH), Baltimore. IEEE*. 2016.

[121] Imad Dabbura. *Predicting Loan Repayment*. https://towardsdatascience.com/predicting-loan-repayment-5df4e0023e92. 2018.

[122] Andreea Danielescu. “Eschewing gender stereotypes in voice assistants to promote inclusion”. In: *Proceedings of the 2nd conference on conversational user interfaces*. 2020, pp. 1–3.

[123] David Danks and Alex John London. “Algorithmic Bias in Autonomous Systems.” In: *IJCAI*. Vol. 17. 2017, pp. 4691–4697.

[124] Rachael A Dansby Olufowote et al. “How can I become more secure?: A grounded theory of earning secure attachment”. In: *Journal of Marital and Family Therapy* 46.3 (2020), pp. 489–506.

[125] Sayamindu Dasgupta and Benjamin Mako Hill. “Scratch community blocks: Supporting children as data scientists”. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM. 2017, pp. 3620–3631.

[126] *Data scientist: A hot job that pays well*. URL: https://www.hiringlab.org/2019/01/17/data-scientist-job-outlook/.

[127] Thomas H Davenport and DJ Patil. *Data scientist: The sexiest job of the 21st Century*. Oct. 2022. URL: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century.

[128] Munmun De Choudhury et al. “Social media participation in an activist movement for racial equality”. In: *Tenth International AAAI Conference on Web and Social Media*. 2016.

[129] Richard D De Veaux et al. “Curriculum guidelines for undergraduate programs in data science”. In: *Annual Review of Statistics and Its Application* 4 (2017), pp. 15–30.

[130] Gerrit De Vynck and Will Oremus. Mar. 2023. URL: https://www.washingtonpost.com/technology/2023/03/30/tech-companies-cut-ai-ethics/.

[131] Erica Deahl. “Better the data you know: Developing youth data literacy in schools and informal learning environments”. In: *Available at SSRN 2445621* (2014).

[132] Richard Delgado. “On telling stories in school: A reply to Farber and Sherry”. In: *Vand. L. Rev.* 46 (1993), p. 665.

[133] Richard Delgado. “Storytelling for oppositionists and others: A plea for narrative”. In: *Michigan law review* 87.8 (1989), pp. 2411–2441.

[134] Richard Delgado. “When a story is just a story: Does voice really matter?” In: *Virginia Law Review* (1990), pp. 95–111.

[135] Richard Delgado and Jean Stefancic. *Critical race theory: An introduction*. Vol. 87. NyU press, 2023.

[136] Michael A DeVito, Darren Gergle, and Jeremy Birnholtz. “" Algorithms ruin everything" # RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media”. In: *Proceedings of the 2017 CHI conference on human factors in computing systems*. 2017, pp. 3163–3174.

[137] Michael Ann DeVito. “Adaptive folk theorization as a path to algorithmic literacy on changing platforms”. In: (2021).

[138] Alicia DeVos et al. “Toward User-Driven Algorithm Auditing: Investigating users strategies for uncovering harmful algorithmic behavior”. en. In: *CHI Conference on Human Factors in Computing Systems*. New Orleans LA USA: ACM, Apr. 2022, pp. 1–19. ISBN: 978-1-4503-9157-3. DOI: 10.1145/3491102.3517441. URL: https://dl.acm.org/doi/10.1145/3491102.3517441 (visited on 08/12/2022).

[139] John Dewey. “Experience and education”. In: *The Educational Forum*. Vol. 50. 3. Taylor & Francis. 1986, pp. 241–252.

[140] Lisa M Diamond. “Contributions of psychophysiology to research on adult attachment: Review and recommendations”. In: *Personality and Social Psychology Review* 5.4 (2001), pp. 276–295.

[141] Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. “Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online”. In: *Sexuality & Culture* 25.2 (2021), pp. 700–732.

[142] William Dieterich, Christina Mendoza, and Tim Brennan. “COMPAS risk scales: Demonstrating accuracy equity and predictive parity”. In: *Northpoint Inc* 7.7.4 (2016), p. 1.

[143] Christina Dinar. *The state of content moderation for the LGBTIQA+ community and the role of the EU Digital Services Act*. Tech. rep. Technical Report. Heinrich-Böll-Stiftung, 2021.

[144] Lydia Dishman. *This is the impact of Instagram’s accidental fat-phobic algorithm*. Oct. 2019. URL: https://www.fastcompany.com/90415917/this-is-the-impact-of-instagrams-accidental-fat-phobic-algorithm.

[145] Ángel Daz and Laura Hecht-Felella. “Double standards in social media content moderation”. In: *Brennan Center for Justice at New York University School of Law. https://www.brennancenter. org/our-work/research-reports/double-standards-socialmedia-content-moderation* (2021).

[146] Amanda Silberling Dominic-Madori Davis. *Metas new AI council is composed entirely of white men*. May 2024. URL: https://techcrunch.com/2024/05/22/metas-new-ai-council-is-comprised-entirely-of-white-men.

[147] David Donoho. “50 years of data science”. In: *Journal of Computational and Graphical Statistics* 26.4 (2017), pp. 745–766.

[148] Benard P Dreyer et al. “The death of George Floyd: bending the arc of history toward justice for generations of children”. In: *Pediatrics* 146.3 (2020).

[149] Stefania Druga et al. “Inclusive AI literacy for kids around the world”. In: *Proceedings of FabLearn 2019*. 2019, pp. 104–111.

[150] Brooke Erin Duffy et al. “The nested precarities of creative labor on social media”. In: *Social Media+ Society* 7.2 (2021), p. 20563051211021368.

[151] Stefanie Duguay, Jean Burgess, and Nicolas Suzor. “Queer womens experiences of patchwork platform governance on Tinder, Instagram, and Vine”. In: *Convergence* 26.2 (2020), pp. 237–252.

[152] Lee Edelman. “No future”. In: *No Future*. Duke University Press, 2004.

[153] Bora Edizel et al. “FaiRecSys: Mitigating algorithmic bias in recommender systems”. In: *International Journal of Data Science and Analytics* 9.2 (2020), pp. 197–213.

[154] Steven D Edwards. “Three versions of an ethics of care”. In: *Nursing philosophy* 10.4 (2009), pp. 231–240.

[155] Johannes C Eichstaedt et al. “The emotional and mental health impact of the murder of George Floyd on the US population”. In: *Proceedings of the National Academy of Sciences* 118.39 (2021), e2109139118.

[156] Elle Hunt. “Tay, Microsoft’s AI chatbot, gets a crash course in racism from Twitter”. In: *The Guardian* (Mar. 2016). URL: https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter?CMP=twt_a-technology_b-gdntech.

[157] Zohar Elyoseph and Inbar Levkovich. “Beyond human expertise: the promise and limitations of ChatGPT in suicide risk assessment”. In: *Frontiers in psychiatry* 14 (2023), p. 1213141.

[158] Motahhare Eslami. “Revisiting Transparency and Fairness in Algorithmic Systems Through the Lens of Public Education and Engagement”. In: *Proceedings of the Eighth ACM Conference on Learning@ Scale*. 2021, pp. 13–13.

[159] Motahhare Eslami et al. “" I always assumed that I wasn’t really that close to [her]" Reasoning about Invisible Algorithms in News Feeds”. In: *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 2015, pp. 153–162.

[160] Motahhare Eslami et al. “Be careful; things can be worse than they appear: Understanding Biased Algorithms and Users Behavior around Them in Rating Platforms”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 11. 1. 2017.

[161] Motahhare Eslami et al. “User attitudes towards algorithmic opacity and transparency in online reviewing platforms”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–14.

[162] Francesco Fabbri et al. “The Effect of Homophily on Disparate Visibility of Minorities in People Recommender Systems”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14. 2020, pp. 165–175.

[163] Hadiya Faheem and Sanjib Dutta. “Artificial Intelligence Failure at IBM’Watson for Oncology’”. In: *IUP Journal of Knowledge Management* 21.3 (2023), pp. 47–75.

[164] Fatemeh Alizadeh and Gunnar Stevens. “Think like a Human, Act like a Bot: Explaining Instagrams Automatic Ban Decisions”. In: *Conference on Human Factors in Computing Systems Extended Abstracts 2020*. Honolulu, Hawaii: ACM, Apr. 2020.

[165] Gretchen Faust. “Hair, Blood and the Nipple. Instagram Censorship and the Female Body”. In: (2017).

[166] RM Pasco Fearon and Glenn I Roisman. “Attachment theory: progress and future directions”. In: *Current Opinion in Psychology* 15 (2017), pp. 131–136.

[167] Claire M Felmingham et al. “The importance of incorporating human factors in the design and implementation of artificial intelligence for skin cancer diagnosis in the real world”. In: *American journal of clinical dermatology* 22.2 (2021), pp. 233–242.

[168] Jessica L Feuston, Alex S Taylor, and Anne Marie Piper. “Conformity of Eating Disorders through Content Moderation”. In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW1 (2020), pp. 1–28.

[169] Casey Fiesler, Natalie Garrett, and Nathan Beard. “What do we teach when we teach tech ethics? A syllabi analysis”. In: *Proceedings of the 51st ACM technical symposium on computer science education*. 2020, pp. 289–295.

[170] Casey Fiesler et al. “Reddit rules! characterizing an ecosystem of governance”. In: *Twelfth International AAAI Conference on Web and Social Media*. 2018.

[171] Ted Fleming. “A Secure Base for Adult Learning: Attachment Theory and Adult Education.” In: *adult learner: the Irish journal of adult and community education* 33 (2008), p. 53.

[172] Luciano Floridi et al. *Ethics, governance, and policies in artificial intelligence*. Springer, 2021.

[173] Luciano Floridi and Josh Cowls. “A unified framework of five principles for AI in society”. In: *Machine learning and the city: Applications in architecture and urban design* (2022), pp. 535–545.

[174] Luciano Floridi and Andrew Strait. “Ethical foresight analysis: what it is and why it is needed?” In: *The 2020 Yearbook of the Digital Ethics Lab* (2021), pp. 173–194.

[175] Luciano Floridi et al. “AI4Peoplean ethical framework for a good AI society: opportunities, risks, principles, and recommendations”. In: *Minds and machines* 28 (2018), pp. 689–707.

[176] Patricia I Fusch Ph D and Lawrence R Ness. “Are we there yet? Data saturation in qualitative research”. In: (2015).

[177] Iason Gabriel. “Toward a theory of justice for artificial intelligence”. In: *Daedalus* 151.2 (2022), pp. 218–231.

[178] Junling Gao et al. “Mental health problems and social media exposure during COVID-19 outbreak”. In: *Plos one* 15.4 (2020), e0231924.

[179] Megan Garcia. “Racist in the machine: The disturbing implications of algorithmic bias”. In: *World Policy Journal* 33.4 (2016), pp. 111–117.

[180] Natalie Garrett, Nathan Beard, and Casey Fiesler. “More than" If Time Allows" the role of ethics in AI education”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020, pp. 272–278.

[181] Marissa Gerchick et al. *How policy hidden in an algorithm is threatening families in this Pennsylvania county: ACLU*. Aug. 2023. URL: https://www.aclu.org/news/womens-rights/how-policy-hidden-in-an-algorithm-is-threatening-families-in-this-pennsylvania-county.

[182] Ysabel Gerrard. “Beyond the hashtag: Circumventing content moderation on social media”. In: *New Media & Society* 20.12 (2018), pp. 4492–4511.

[183] Ysabel Gerrard. “Social media content moderation: Six opportunities for feminist intervention”. In: *Feminist Media Studies* 20.5 (2020), pp. 748–751.

[184] Ysabel Gerrard and Helen Thornham. “Content moderation: Social medias sexist assemblages”. In: *new media & society* 22.7 (2020), pp. 1266–1286.

[185] Tarleton Gillespie. “Content moderation, AI, and the question of scale”. In: *Big Data & Society* 7.2 (2020), p. 2053951720943234.

[186] Tarleton Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

[187] Carol Gilligan. *In a different voice: Psychological theory and womens development*. Harvard university press, 1993.

[188] Jessica L Gillotte. “Copyright infringement in ai-generated artworks”. In: *UC Davis L. Rev.* 53 (2019), p. 2655.

[189] Evelyn Nakano Glenn. “The social construction and institutionalization of gender and race: An integrative framework”. In: *Revisioning gender* (1999), pp. 3–43.

[190] Trystan S Goetze. "AI Art is Theft: Labour, Extraction, and Exploitation, Or, On the Dangers of Stochastic Pollocks". In: *arXiv preprint arXiv:2401.06178* (2024).

[191] Judy Goldsmith and Emanuelle Burton. "Why teaching ethics to AI practitioners is important". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 1. 2017.

[192] Norma Gonzalez et al. "Funds of knowledge for teaching in Latino households". In: *Urban Education* 29.4 (1995), pp. 443–470.

[193] Norma González, Luis C Moll, and Cathy Amanti. *Funds of knowledge: Theorizing practices in households, communities, and classrooms*. Routledge, 2006.

[194] Dan Goodley. "Empowerment, self-advocacy and resilience". In: *Journal of Intellectual Disabilities* 9.4 (2005), pp. 333–343.

[195] Dan Goodley. *Self-advocacy in the lives of people with learning difficulties: The politics of resilience*. Open University Press Buckingham, 2000.

[196] Isabel Goodwin. "The relevance of attachment theory to the philosophy, organization, and practice of adult mental health care". In: *Clinical Psychology Review* 23.1 (2003), pp. 35–56.

[197] Robert Gorwa, Reuben Binns, and Christian Katzenbach. "Algorithmic content moderation: Technical and political challenges in the automation of platform governance". In: *Big Data & Society* 7.1 (2020), p. 2053951719897945.

[198] Nico Grant and Kashmir Hill. *Googles photo app still cant find gorillas. and neither can Apples.* May 2023. URL: https://www.nytimes.com/2023/05/22/technology/ai-photo-labels-google-apple.html.

[199] Mary L Gray and Siddharth Suri. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books, 2019.

[200] Jami Grich. *Earned secure attachment in young adulthood: Adjustment and relationship satisfaction*. Texas A&M University, 2001.

[201] Nicole Gross. "What chatGPT tells us about gender: a cautionary tale about performativity and gender biases in AI". In: *Social Sciences* 12.8 (2023), p. 435.

[202] Douglas Guilbeault et al. "Online images amplify gender bias". In: *Nature* (2024), pp. 1–7.

[203] Kristin K Gundersen and Kristen L Zaleski. "Posting the story of your sexual assault online: A phenomenological study of the aftermath". In: *Feminist Media Studies* 21.5 (2021), pp. 840–852.

[204] Lisa N Guo et al. "Bias in, bias out: underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detectiona scoping review". In: *Journal of the American Academy of Dermatology* 87.1 (2022), pp. 157–159.

[205] Maanak Gupta et al. "From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy". In: *IEEE Access* (2023).

[206] Edwin Ray Guthrie. "Conditioning as a principle of learning." In: *Psychological review* 37.5 (1930), p. 412.

[207] Jessica Guynn. *Google photos labeled black people gorillas*. July 2015. URL: https://www.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-as-gorillas/29567465/.

[208] Thilo Hagendorff. "AI ethics and its pitfalls: not living up to its own standards?" In: *AI and Ethics* 3.1 (2023), pp. 329–336.

[209] Thilo Hagendorff. “AI virtues–The missing link in putting AI ethics into practice”. In: *arXiv preprint arXiv:2011.12750* (2020).

[210] Thilo Hagendorff. “The ethics of AI ethics: An evaluation of guidelines”. In: *Minds and machines* 30.1 (2020), pp. 99–120.

[211] Hani Hagras. “Toward human-understandable, explainable AI”. In: *Computer* 51.9 (2018), pp. 28–36.

[212] Oliver L Haimson et al. “Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas”. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (2021), pp. 1–35.

[213] Kori Hale. *A.I. bias caused 80 percent of black mortgage applicants to be denied*. Nov. 2022. URL: https://www.forbes.com/sites/korihale/2021/09/02/ai-bias-caused-80-of-black-mortgage-applicants-to-be-denied/?sh=268b796536fe.

[214] Donna Haraway. “Situated knowledges: The science question in feminism and the privilege of partial perspective”. In: *Feminist studies* 14.3 (1988), pp. 575–599.

[215] Donna Haraway. “Situated knowledges: The science question in feminism and the privilege of partial perspective”. In: *Feminist theory reader*. Routledge, 2020, pp. 303–310.

[216] Johanna Hardin et al. “Data science in statistics curricula: Preparing students to think with data”. In: *The American Statistician* 69.4 (2015), pp. 343–353.

[217] Idit Ed Harel and Seymour Ed Papert. *Constructionism.* Ablex Publishing, 1991.

[218] Elizabeth Harlow. “Attachment theory: Developments, debates and recent applications in social work, social care and education”. In: *Journal of Social Work Practice* 35.1 (2021), pp. 79–91.

[219] Andrea Hartzler and Wanda Pratt. “Managing the personal side of health: how patient expertise differs from the expertise of clinicians”. In: *Journal of medical Internet research* 13.3 (2011), e62.

[220] Bertrand K Hassani. “Societal bias reinforcement through machine learning: a credit scoring perspective”. In: *AI and Ethics* 1.3 (2021), pp. 239–247.

[221] Janna Hastings. “Preventing harm from non-conscious bias in medical generative AI”. In: *The Lancet Digital Health* 6.1 (2024), e2–e3.

[222] Samantha Hautea, Sayamindu Dasgupta, and Benjamin Mako Hill. “Youth perspectives on critical data literacies”. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2017, pp. 919–930.

[223] Deion S Hawkins. “After Philando, I had to take a sick day to recover: Psychological distress, trauma and police brutality in the Black community”. In: *Health communication* 37.9 (2022), pp. 1113–1122.

[224] H Carl Haywood. “What is cognitive education? The view from 30,000 feet”. In: *Journal of Cognitive Education and Psychology* 12.1 (2013), pp. 26–44.

[225] Birte Heinemann et al. “Drafting a data science curriculum for secondary schools”. In: *Proceedings of the 18th Koli Calling International Conference on Computing Education Research*. 2018, pp. 1–5.

[226] Hanneke Hendriks et al. “Picture me drinking: alcohol-related posts by Instagram influencers popular among adolescents and young adults”. In: *Frontiers in psychology* 10 (2020), p. 2991.

[227] Joseph Henrich, Steven J Heine, and Ara Norenzayan. “Most people are not WEIRD”. In: *Nature* 466.7302 (2010), pp. 29–29.

[228] Eelco Herder and Boping Zhang. “Unexpected and unpredictable: Factors that make personalized advertisements creepy”. In: *Proceedings of the 23rd International Workshop on Personalization and Recommendation on the Web and Beyond*. 2019, pp. 1–6.

[229] Stephanie C Hicks and Rafael A Irizarry. “A guide to teaching data science”. In: *The American Statistician* 72.4 (2018), pp. 382–391.

[230] K Hill. “Accused of cheating by an algorithm, and a professor she had never met”. In: *New York Times* 27 (2022).

[231] Kashmir Hill. “Wrongfully accused by an algorithm”. In: *Ethics of Data and Analytics*. Auerbach Publications, 2022, pp. 138–142.

[232] Alexis Hiniker et al. “MyTime: designing and evaluating an intervention for smartphone non-use”. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2016, pp. 4746–4757.

[233] Robert Hinson et al. “Antecedents and consequences of customer engagement on Facebook: An attachment theory perspective”. In: *Journal of Research in Interactive Marketing* (2019).

[234] Tom Hitron et al. “Can Children Understand Machine Learning Concepts?: The Effect of Uncovering Black Boxes”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM. 2019, p. 415.

[235] Fred Hohman et al. “Gamut: A design probe to understand how data scientists understand machine learning models”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM. 2019, p. 579.

[236] Kenneth Holstein et al. “Improving fairness in machine learning systems: What do industry practitioners need?” In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM. 2019, p. 600.

[237] A. Holzinger. “From Machine Learning to Explainable AI”. In: *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*. 2018, pp. 55–66. DOI: 10.1109/DISA.2018.8490530.

[238] Md Mahbub Hossain et al. “Epidemiology of mental health problems in COVID-19: a review”. In: *F1000Research* 9 (2020).

[239] David Howe. *Attachment theory for social work practice*. Macmillan International Higher Education, 1995.

[240] Hubspot. *Instagram Marketing*. 2022. URL: https://www.hubspot.com/instagram-marketing.

[241] Eslam Hussein, Prerna Juneja, and Tanushree Mitra. “Measuring misinformation in video search platforms: An audit study on YouTube”. In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW1 (2020), pp. 1–27.

[242] Jane Im et al. “Yes: Affirmative consent as a theoretical framework for understanding and imagining social platforms”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–18.

[243] Instagram. *See Posts you Care About First in your Feed*. Mar. 2016. URL: https://about.instagram.com/blog/announcements/see-posts-you-care-about-first-in-your-feed.

[244] *Is data scientist still the sexiest job of the 21st century?* July 2022. URL: https://hbr.org/2022/07/is-data-scientist-still-the-sexiest-job-of-the-21st-century.

[245] Isidora Janezic and Stephanie Arsenault. "How to foster empathy in anti-discrimination initiatives? Implication for social interventionsa qualitative approach". In: *Canadian Ethnic Studies* 53.1 (2021), pp. 47–68.

[246] Jeffrey Dastin. "Amazon scraps secret AI recruiting tool that showed bias against women". In: *Reuters* (Oct. 2018). URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

[247] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. "Does transparency in moderation really matter? User behavior after content removal explanations on reddit". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019), pp. 1–27.

[248] Shagun Jhaver, Yoni Karpfen, and Judd Antin. "Algorithmic Anxiety and Coping Strategies of Airbnb Hosts". en. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Montreal QC Canada: ACM, Apr. 2018, pp. 1–12. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3173995. URL: https://dl.acm.org/doi/10.1145/3173574.3173995 (visited on 12/05/2022).

[249] Shagun Jhaver et al. "" Did you suspect the post would be removed?" Understanding user reactions to content removals on Reddit". In: *Proceedings of the ACM on human-computer interaction* 3.CSCW (2019), pp. 1–33.

[250] Shagun Jhaver et al. ""Did You Suspect the Post Would be Removed?": Understanding User Reactions to Content Removals on Reddit". en. In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (Nov. 2019), pp. 1–33. ISSN: 2573-0142. DOI: 10.1145/3359294. URL: https://dl.acm.org/doi/10.1145/3359294 (visited on 12/12/2022).

[251] Anna Jobin, Marcello Ienca, and Effy Vayena. "The global landscape of AI ethics guidelines". In: *Nature machine intelligence* 1.9 (2019), pp. 389–399.

[252] Gabbrielle M Johnson. "Algorithmic bias: on the implicit biases of social technology". In: *Synthese* 198.10 (2021), pp. 9941–9961.

[253] Javon Johnson. "Black joy in the time of Ferguson". In: *QED: A Journal in GLBTQ Worldmaking* 2.2 (2015), pp. 177–183.

[254] Thaddeus L Johnson et al. "Facial recognition systems in policing and racial disparities in arrests". In: *Government Information Quarterly* 39.4 (2022), p. 101753.

[255] David H Jonassen, Jamie M Myers, and Ann Margaret McKillop. "From constructivism to constructionism: Learning with hypermedia/multimedia rather than from it". In: *Constructivist learning environments: Case studies in instructional design* (1996), pp. 93–106.

[256] Paul Joseph-Richard and James Uhomoibhi. "Which Data Sets Are Preferred by University Students in Learning Analytics Dashboards? A Situated Learning Theory Perspective". In: *INFORMS Transactions on Education* (2023).

[257] Maja Kabiljo and Aleksandar Ilic. "Recommending items to more than a billion people". In: *Retrieved May* 2 (2015), p. 2018.

[258] Kari Paul. "Senate panel interrogates Instagram CEO on how platform protects children". In: *The Guardian* (Dec. 2021). URL: https://www.theguardian.com/technology/2021/dec/08/instagram-adam-mosseri-us-congress-testimony.

[259] Nadia Karizat et al. “Algorithmic Folk Theories and Identity: How TikTok Users Co-Produce Knowledge of Identity and Engage in Algorithmic Resistance”. In: (2021).

[260] Nadia Karizat et al. “Algorithmic Folk Theories and Identity: How TikTok Users Co-Produce Knowledge of Identity and Engage in Algorithmic Resistance”. en. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (Oct. 2021), pp. 1–44. ISSN: 2573-0142. DOI: 10.1145/3476046. URL: https://dl.acm.org/doi/10.1145/3476046 (visited on 08/12/2022).

[261] Nicolas Kayser-Bril. *LinkedIn automatically rates out-of-country candidates as not fit in job applications*. URL: https://algorithmwatch.org/en/linkedin-recruitment-feature-discrimination/.

[262] Ibram X Kendi. *How to be an antiracist*. One world, 2023.

[263] Ian Kennedy et al. “Repeat Spreaders and Election Delegitimization: A Comprehensive Dataset of Misinformation Tweets from the 2020 US Election”. In: *Journal of Quantitative Description: Digital Media* 2 (2022).

[264] C. D. Kidd and C. Breazeal. “Effect of a robot on user perceptions”. In: *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*. Vol. 4. Sept. 2004, 3559–3564 vol.4. DOI: 10.1109/IROS.2004.1389967.

[265] Celeste Kidd, Holly Palmeri, and Richard N Aslin. “Rational snacking: Young childrens decision-making on the marshmallow task is moderated by beliefs about environmental reliability”. In: *Cognition* 126.1 (2013), pp. 109–114.

[266] Eunhyang Kim and Eunyoung Koh. “Avoidant attachment and smartphone addiction in college students: The mediating effects of anxiety and self-esteem”. In: *Computers in Human Behavior* 84 (2018), pp. 264–271.

[267] Nam Wook Kim et al. “DataSelfie: Empowering People to Design Personalized Visuals to Represent Their Data”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM. 2019, p. 79.

[268] Ezekiel W Kimball et al. “College students with disabilities redefine activism: Self-advocacy, storytelling, and collective action.” In: *Journal of Diversity in Higher Education* 9.3 (2016), p. 245.

[269] Sara Kingsley et al. “Auditing Digital Platforms for Discrimination in Economic Opportunity Advertising”. In: *arXiv preprint arXiv:2008.09656* (2020).

[270] Jon Kleinberg et al. “Human decisions and machine predictions”. In: *The quarterly journal of economics* 133.1 (2017), pp. 237–293.

[271] Rob Knapp. “Wholesome design for wicked problems”. In: *The Public Sphere Project. http://www. publicsphereproject. org/content/wholesome-design-wicked-problems. Zugegriffen* 15 (2008).

[272] Thomas David Knestrict. *A post-modern critique of attachment theory: Moving towards a socially just ecological framework*. University of Cincinnati, 2002.

[273] Bart P Knijnenburg and Shlomo Berkovsky. “Privacy for recommender systems: tutorial abstract”. In: *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 2017, pp. 394–395.

[274] Amy J Ko et al. “Critically conscious computing: Methods for secondary education”. In: *CONSULTATO: 10* (2021).

[275] Amy J Ko et al. “It is time for more critical CS education”. In: *Communications of the ACM* 63.11 (2020), pp. 31–33.

[276] Allison Koenecke et al. “Racial disparities in automated speech recognition”. In: *Proceedings of the National Academy of Sciences* 117.14 (2020), pp. 7684–7689.

[277] Kolina Koltai, Rachel E Moran, and Izzi Grasso. “Addressing the root of vaccine hesitancy during the COVID-19 pandemic”. In: *XRDS: Crossroads, The ACM Magazine for Students* 28.2 (2022), pp. 34–38.

[278] Vladimir J Koneni. “Responsible behavioral science generalizations and applications require much more than non-WEIRD samples”. In: *Behavioral and Brain Sciences* 33.2-3 (2010), pp. 98–99.

[279] Konstantina Kourou et al. “Machine learning applications in cancer prognosis and prediction”. In: *Computational and structural biotechnology journal* 13 (2015), pp. 8–17.

[280] Sean Kross and Philip J Guo. “Practitioners Teaching Data Science in Industry and Academia: Expectations, Workflows, and Challenges”. In: (2019).

[281] James Kuczmarski. *Reducing gender bias in Google Translate*. Dec. 2018. URL: https://blog.google/products/translate/reducing-gender-bias-google-translate/.

[282] Paul B de Laat. “Companies committed to responsible AI: From principles towards implementation and regulation?” In: *Philosophy & technology* 34 (2021), pp. 1135–1193.

[283] Nancy K Lankton and D Harrison McKnight. “What does it mean to trust facebook? Examining technology and interpersonal trust beliefs”. In: *ACM SIGMIS Database: the DATABASE for Advances in Information Systems* 42.2 (2011), pp. 32–54.

[284] Jeff Larson et al. *How we analyzed the compas recidivism algorithm*. May 2016. URL: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[285] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. “Alexa, are you listening? Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers”. In: *Proceedings of the ACM on human-computer interaction* 2.CSCW (2018), pp. 1–31.

[286] Jean Lave, Etienne Wenger, et al. *Situated learning: Legitimate peripheral participation*. Cambridge university press, 1991.

[287] Vinhcent Le. Dec. 2022. URL: https://greenlining.org/publications/algorithmic-bias-explained/.

[288] Min Kyung Lee et al. “Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers”. en. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Seoul Republic of Korea: ACM, Apr. 2015, pp. 1603–1612. ISBN: 978-1-4503-3145-6. DOI: 10.1145/2702123.2702548. URL: https://dl.acm.org/doi/10.1145/2702123.2702548 (visited on 12/05/2022).

[289] Nicol Turner Lee, Paul Resnick, and Genie Barton. “Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms”. In: *Brookings Institute: Washington, DC, USA* (2019).

[290] Pengfei Li et al. “Making AI Less" Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models”. In: *arXiv preprint arXiv:2304.03271* (2023).

[291] Calvin A Liang, Sean A Munson, and Julie A Kientz. “Embracing four tensions in human-computer interaction research with marginalized people”. In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 28.2 (2021), pp. 1–47.

[292] Anne-Laure Ligozat et al. “Unraveling the hidden environmental impacts of AI solutions for environment life cycle assessment of AI solutions”. In: *Sustainability* 14.9 (2022), p. 5172.

[293] Lanna Lima et al. “Empirical analysis of bias in voice-based personal assistants”. In: *Companion Proceedings of the 2019 World Wide Web Conference*. 2019, pp. 533–538.

[294] Emma J Llansó. “No amount of AI in content moderation will solve filterings prior-restraint problem”. In: *Big Data & Society* 7.1 (2020), p. 2053951720920686.

[295] Duri Long and Brian Magerko. “What is AI literacy? Competencies and design considerations”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–16.

[296] Mufan Luo and Jeffrey T Hancock. “Self-disclosure and social media: motivations, mechanisms and psychological well-being”. In: *Current Opinion in Psychology* 31 (2020), pp. 110–115.

[297] Deborah Lupton. “You are your data: Self-tracking practices and concepts of data”. In: *Lifelogging*. Springer, 2016, pp. 61–79.

[298] Renkai Ma and Yubo Kou. “" How advertiser-friendly is my video?": YouTuber’s Socioeconomic Interactions with Algorithmic Content Moderation”. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (2021), pp. 1–25.

[299] Sean MacAvaney et al. “Hate speech detection: Challenges and solutions”. In: *PloS one* 14.8 (2019), e0221152.

[300] Kelly Mack et al. “What do we mean by accessibility research? A literature survey of accessibility papers in CHI and ASSETS from 1994 to 2019”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–18.

[301] João Carlos Magalhães and Christian Katzenbach. “Coronavirus and the frailness of platform governance”. In: *Internet Policy Review* 9 (2020).

[302] Mary Main, Judith Solomon, et al. “Procedures for identifying infants as disorganized/disoriented during the Ainsworth Strange Situation”. In: *Attachment in the preschool years: Theory, research, and intervention* 1 (1990), pp. 121–160.

[303] Shervin Malmasi and Marcos Zampieri. “Detecting hate speech in social media”. In: *arXiv preprint arXiv:1712.06427* (2017).

[304] James Manyika et al. “Big data: The next frontier for innovation, competition, and productivity”. In: (2011).

[305] Brandeis Marshall. *Algorithmic misogynoir in content moderation practice*. Tech. rep. Technical Report. Heinrich-Böll-Stiftung, 2021.

[306] Aja Y Martinez. “A plea for critical race theory counterstory: Stock story versus counterstory dialogues concerning Alejandra’s" fit" in the academy”. In: *Composition Studies* (2014), pp. 33–55.

[307] Emmanuel Martinez and Lauren Kirchner. “The secret bias hidden in mortgage-approval algorithms”. In: *The Markup* (2021).

[308] Alex Marzano-Lesnevich. “Flying While Trans.” In: *International New York Times* (2019), NA–NA.

[309] Camillia Matuk et al. “Data Literacy for Social Justice”. In: ().

[310] Sandra G Mayson. “Bias in, bias out”. In: *YAle lJ* 128 (2018), p. 2218.

[311] Ally McCrow-Young. “Approaching Instagram data: reflections on accessing, archiving and anonymising visual social media”. In: *Communication Research and Practice* 7.1 (2021), pp. 21–34.

[312] Brad McKenna and Hameed Chughtai. “Resistance and sexuality in virtual worlds: An LGBT perspective”. In: *Computers in Human Behavior* 105 (2020), p. 106199.

[313] Sian A McLean et al. “A pilot evaluation of a social media literacy intervention to reduce risk factors for eating disorders”. In: *International Journal of Eating Disorders* 50.7 (2017), pp. 847–851.

[314] Andrew McNamara, Justin Smith, and Emerson Murphy-Hill. “Does ACMs code of ethics change ethical decision making in software development?” In: *Proceedings of the 2018 26th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering*. 2018, pp. 729–733.

[315] Lachlan A McWilliams and Gordon JG Asmundson. “The relationship of adult attachment dimensions to pain-related fear, hypervigilance, and catastrophizing”. In: *Pain* 127.1-2 (2007), pp. 27–34.

[316] Ninareh Mehrabi et al. “A survey on bias and fairness in machine learning”. In: *ACM computing surveys (CSUR)* 54.6 (2021), pp. 1–35.

[317] Paul Mena, Danielle Barbe, and Sylvia Chan-Olmsted. “Misinformation on Instagram: The impact of trusted endorsements on message credibility”. In: *Social Media+ Society* 6.2 (2020), p. 2056305120935102.

[318] Zion Mengesha et al. “I dont Think These Devices are Very Culturally Sensitive.Impact of Automated Speech Recognition Errors on African Americans”. In: *Frontiers in Artificial Intelligence* 4 (2021), p. 169.

[319] Lisa R Merriweather Hunn, Talmadge C Guy, and Elaine Mangliitz. “Who can speak for whom? Using counter-storytelling to challenge racial hegemony”. In: (2006).

[320] Fabien Merz. “Europe and the global AI race”. In: *CSS analyses in security policy* 247 (2019).

[321] Thaddeus Metz and Sarah Clark Miller. “Relational ethics”. In: (2016).

[322] Richard Miller, Katrina Liu, and Arnetha F Ball. “Critical counter-narrative as transformative methodology for educational equity”. In: *Review of Research in Education* 44.1 (2020), pp. 269–300.

[323] Ryan A Miller and Annemarie Vaccaro. “Queer student leaders of color: Leadership as authentic, collaborative, culturally competent”. In: *Journal of Student Affairs Research and Practice* 53.1 (2016), pp. 39–50.

[324] Steven Miller and Debbie Hughes. “The quant crunch: How the demand for data science skills is disrupting the job market”. In: *Burning Glass Technologies* (2017).

[325] Dan Milmo and Alex Hern. *we definitely messed up: Why did google ai tool make offensive historical images?* Mar. 2024. URL: https://www.theguardian.com/technology/2024/mar/08/we-definitely-messed-up-why-did-google-ai-tool-make-offensive-historical-images.

[326] Margaret Mitchell et al. “Model cards for model reporting”. In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 220–229.

[327] Brent Mittelstadt. “Principles alone cannot guarantee ethical AI”. In: *Nature machine intelligence* 1.11 (2019), pp. 501–507.

[328] Tamar Mitts, Nilima Pisharody, and Jacob Shapiro. "Removal of Anti-Vaccine Content Impacts Social Media Discourse". In: *14th ACM Web Science Conference 2022*. 2022, pp. 319–326.

[329] Luis C Moll et al. "Funds of knowledge for teaching: Using a qualitative approach to connect homes and classrooms". In: *Theory into practice* 31.2 (1992), pp. 132–141.

[330] Megan A Moreno and Jennifer M Whitehill. "Influence of social media on alcohol use in adolescents and young adults". In: *Alcohol research: current reviews* 36.1 (2014), p. 91.

[331] Jessica Morley et al. "Ethics as a service: a pragmatic operationalisation of AI ethics". In: *Minds and Machines* 31.2 (2021), pp. 239–256.

[332] Jessica Morley et al. "From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices". In: *Science and engineering ethics* 26.4 (2020), pp. 2141–2168.

[333] Jessica Morley et al. "Operationalising AI ethics: barriers, enablers and next steps". In: *AI & SOCIETY* (2021), pp. 1–13.

[334] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. "Hate speech detection and racial bias mitigation in social media based on BERT model". In: *PloS one* 15.8 (2020), e0237861.

[335] Robert T Muller. "Trauma and dismissing (avoidant) attachment: Intervention strategies in individual psychotherapy." In: *Psychotherapy: Theory, Research, Practice, Training* 46.1 (2009), p. 68.

[336] Luke Munn. "The uselessness of AI ethics". In: *AI and Ethics* (2022), pp. 1–9.

[337] Evelyne Musambi. *Facebooks parent company Meta and moderators suing it for* $1.6 billion in Kenya agr$ Aug. 2023. URL: https://apnews.com/article/kenya-facebook-content-moderators-meta-lawsuit-sama-5dca81fa5df9aa87886366945818dfa9.

[338] Tyler Musgrave, Alia Cummings, and Sarita Schoenebeck. "Experiences of Harm, Healing, and Joy among Black Women and Femmes on Social Media". en. In: *CHI Conference on Human Factors in Computing Systems*. New Orleans LA USA: ACM, Apr. 2022, pp. 1–17. ISBN: 978-1-4503-9157-3. DOI: 10.1145/3491102.3517608. URL: https://dl.acm.org/doi/10.1145/3491102.3517608 (visited on 08/19/2022).

[339] Tyler Musgrave, Alia Cummings, and Sarita Schoenebeck. "Experiences of Harm, Healing, and Joy among Black Women and Femmes on Social Media". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3517608. URL: https://doi.org/10.1145/3491102.3517608.

[340] Sarah Myers West. "Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms". In: *New Media & Society* 20.11 (2018), pp. 4366–4383.

[341] Yifat Nahmias and Maayan Perel. "The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations". In: *Harv. J. on Legis.* 58 (2021), p. 145.

[342] *Nala Cat*. URL: https://www.instagram.com/nala_cat/?hl=en.

[343] Engineering National Academies of Sciences, Medicine, et al. *How people learn II: Learners, contexts, and cultures*. National Academies Press, 2018.

[344] Peter Naur. "Concise survey of computer methods". In: *(No Title)* (1974).

[345] Rachel Naylor. *Facebook trial lets users hide alcohol adverts*. Sept. 2017. URL: https://www.bbc.com/news/technology-41332082.

[346] Todd W Neller et al. "Model AI Assignments 2022". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 11. 2022, pp. 12863–12864.

[347] Greg L Nelson, Benjamin Xie, and Amy J Ko. "Comprehension first: evaluating a novel pedagogy and tutoring system for program tracing in CS1". In: *Proceedings of the 2017 ACM Conference on International Computing Education Research*. ACM. 2017, pp. 2–11.

[348] Paul Nemitz. "Constitutional democracy and technology in the age of artificial intelligence". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376.2133 (2018), p. 20180089.

[349] Andrew Ng. *Machine Learning Coursera Course*. https://www.coursera.org/learn/machine-learning. 2011.

[350] James Nicholls. "Everyday, everywhere: alcohol marketing and social mediacurrent trends". In: *Alcohol and alcoholism* 47.4 (2012), pp. 486–493.

[351] Leonardo Nicoletti and Dina Bass. *Humans are biased. Generative AI is even worse*. 2023. URL: https://www.bloomberg.com/graphics/2023-generative-ai-bias/.

[352] Safiya Umoja Noble. *Algorithms of oppression: How search engines reinforce racism*. nyu Press, 2018.

[353] Nel Noddings. "The caring relation in teaching". In: *Oxford review of education* 38.6 (2012), pp. 771–781.

[354] Nel Noddings. "The language of care ethics". In: *Knowledge Quest* 40.5 (2012), p. 52.

[355] R Nonomura et al. "Toward a trauma-and violence-informed research ethics module: Considerations and recommendations". In: *London, ON: Centre for Research & Education on Violence Against Women & Children, Western University* (2020).

[356] Emil Noordeh et al. "Echo Chambers in Collaborative Filtering Based Recommendation Systems". In: *arXiv preprint arXiv:2011.03890* (2020).

[357] Illah Reza Nourbakhsh. "AI ethics: a call to faculty". In: *Communications of the ACM* 64.9 (2021), pp. 43–45.

[358] Lorelli S Nowell et al. "Thematic analysis: Striving to meet the trustworthiness criteria". In: *International journal of qualitative methods* 16.1 (2017), p. 1609406917733847.

[359] Kendall O'Brien. "The cultivation of eating disorders through Instagram". In: (2015).

[360] Cathy O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2016.

[361] Derek OCallaghan et al. "Down the (white) rabbit hole: The extreme right and online recommender systems". In: *Social Science Computer Review* 33.4 (2015), pp. 459–478.

[362] Ziad Obermeyer et al. "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464 (2019), pp. 447–453.

[363] Ihudiya Finda Ogbonnaya-Ogburu et al. "Critical race theory for HCI". In: *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020, pp. 1–16.

[364] Sofia Olhede and Russell Rodrigues. "Fairness and transparency in the age of the algorithm". In: *Significance* 14.2 (2017), pp. 8–9.

[365] Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. “Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online”. In: *Sexuality & Culture* (2020), pp. 1–33.

[366] Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. “Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online”. In: *Sexuality & Culture* 25.2 (2021), pp. 700–732.

[367] Jeffrey C Oliver and Torbet McNeil. “Undergraduate data science degrees emphasize computer science and statistics but fall short in ethics training and domain-specific context”. In: *PeerJ Computer Science* 7 (2021), e441.

[368] Eric D Olson and Kelly Reddy-Best. “Pre-topsurgery, the body scanning machine would most likely error: Transgender and gender nonconforming travel and tourism experiences”. In: *Tourism Management* 70 (2019), pp. 250–261.

[369] Alexandra Olteanu, Kartik Talamadupula, and Kush R Varshney. “The limits of abstract evaluation metrics: The case of hate speech detection”. In: *Proceedings of the 2017 ACM on web science conference*. 2017, pp. 405–406.

[370] Ramona L Paetzold, W Steven Rholes, and Jamie L Kohn. “Disorganized attachment in adulthood: Theory, measurement, and implications for romantic relationships”. In: *Review of General Psychology* 19.2 (2015), pp. 146–156.

[371] Michael Palmer. “Data is the new oil”. In: *ANA marketing maestros* 3 (2006).

[372] Trishan Panch, Heather Mattie, and Rifat Atun. “Artificial intelligence and algorithmic bias: implications for health systems”. In: *Journal of global health* 9.2 (2019).

[373] Chiara Panciroli et al. “Towards AI literacy: A proposal of a framework based on the Episodes of Situated Learning”. In: (2022).

[374] Seymour Papert. *Mindstorms: Children, computers, and powerful ideas*. Basic Books, Inc., 1980.

[375] Seymour Papert. “Papert on piaget”. In: *Time magazine, pág* 105 (1999).

[376] Seymour Papert. “The children’s machine: Rethinking school in the age of the computer”. In: *New York* (1993).

[377] Seymour Papert and Idit Harel. “Situating constructionism”. In: *constructionism* 36.2 (1991), pp. 1–11.

[378] Laurie Anne Pearlman and Christine A Courtois. “Clinical applications of the attachment framework: Relational treatment of complex trauma”. In: *Journal of Traumatic Stress: Official Publication of The International Society for Traumatic Stress Studies* 18.5 (2005), pp. 449–459.

[379] Jane L Pearson et al. “Earned-and continuous-security in adult attachment: Relation to depressive symptomatology and parenting style”. In: *Development and psychopathology* 6.2 (1994), pp. 359–373.

[380] Evan M Peck, Sofia E Ayuso, and Omar El-Etr. “Data is Personal: Attitudes and Perceptions of Data Visualization in Rural Pennsylvania”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM. 2019, p. 244.

[381] Sidney Perkowitz. “The bias in the machine: Facial recognition technology and racial disparities”. In: (2021).

[382] Billy Perrigo. *Inside Facebooks African sweatshop*. Feb. 2022. URL: `https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/`.

[383] Caitlin Petre, Brooke Erin Duffy, and Emily Hund. "Gaming the system: Platform paternalism and the politics of algorithmic visibility". In: *Social Media+ Society* 5.4 (2019), p. 2056305119879995.

[384] Anne Pfeifle. "Alexa, what should we do about privacy: Protecting privacy for users of voice-activated devices". In: *Wash. L. Rev.* 93 (2018), p. 421.

[385] Michael S Pollard, Joan S Tucker, and Harold D Green. "Changes in adult alcohol use and consequences during the COVID-19 pandemic in the US". In: *JAMA network open* 3.9 (2020), e2022942–e2022942.

[386] Michael I Posner and Mary K Rothbart. "Developing mechanisms of self-regulation". In: *Development and psychopathology* 12.3 (2000), pp. 427–441.

[387] Javier Calzada Prado and Miguel Ángel Marzal. "Incorporating data literacy into information literacy programs: Core competencies and contents". In: *Libri* 63.2 (2013), pp. 123–134.

[388] A Treatment Improvement Protocol. "Trauma-informed care in behavioral health services". In: *Rockville, USA: Substance Abuse and Mental Health Services Administration* (2014).

[389] Pushshift. *Pushshift*. 2019. URL: `https://github.com/pushshift/api`.

[390] Walter Quattrociocchi, Antonio Scala, and Cass R Sunstein. "Echo chambers on Facebook". In: *Available at SSRN 2795110* (2016).

[391] Emma K Quinn, Sajjad S Fazel, and Cheryl E Peters. "The Instagram infodemic: cobranding of conspiracy theories, coronavirus disease 2019 and authority-questioning beliefs". In: *Cyberpsychology, Behavior, and Social Networking* 24.8 (2021), pp. 573–577.

[392] Sadruddin Bahadur Qutoshi. "Phenomenology: A philosophy and method of inquiry." In: *Journal of Education and Educational Development* 5.1 (2018), pp. 215–222.

[393] Emilee Rader and Rebecca Gray. "Understanding user beliefs about algorithmic curation in the Facebook news feed". In: *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 2015, pp. 173–182.

[394] Inioluwa Deborah Raji and Joy Buolamwini. "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019, pp. 429–435.

[395] Inioluwa Deborah Raji, Morgan Klaus Scheuerman, and Razvan Amironesei. "You can't sit with us: exclusionary pedagogy in AI ethics education". In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021, pp. 515–525.

[396] Inioluwa Deborah Raji et al. "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing". In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020, pp. 33–44.

[397] Yim Register and Amy J. Ko. "Learning Machine Learning with Personal Data Helps Stakeholders Ground Advocacy Arguments in Model Mechanics". In: *Proceedings of the 2020 ACM Conference on International Computing Education Research*. ICER 20. Virtual Event, New Zealand: Association for Computing Machinery, 2020, pp. 67–78.

[398] Yim Register and Emma S Spiro. "Developing Self-Advocacy Skills through Machine Learning Education: The Case of Ad Recommendation on Facebook". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 16. 2022, pp. 817–828.

[399] Yim Register et al. “AI education matters: Guiding our Future AI Leaders with Joy and Justice”. In: *AI Matters* 8.2 (2022), pp. 22–24.

[400] Yim Register et al. “Attached to The Algorithm: Making Sense of Algorithmic Precarity on Instagram”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–15.

[401] Yim Register et al. “Beyond Initial Removal: Lasting Impacts of Discriminatory Content Moderation to Marginalized Creators on Instagram”. In: *Computer Supported Cooperative Work (CSCW)* 8 (2024).

[402] Mitchel Resnick et al. “Scratch: programming for all”. In: *Communications of the ACM* 52.11 (2009), pp. 60–67.

[403] Anas Rességuier and Rowena Rodrigues. “AI ethics should not remain toothless! A call to bring back the teeth of ethics”. In: *Big Data & Society* 7.2 (2020), p. 2053951720942541.

[404] Horst WJ Rittel and Melvin M Webber. “Dilemmas in a general theory of planning”. In: *Policy sciences* 4.2 (1973), pp. 155–169.

[405] Sarah T Roberts. *Behind the screen*. Yale University Press, 2019.

[406] Lauren Rouse and Anastasia Salter. “Cosplay on demand? Instagram, OnlyFans, and the gendered fantrepreneur”. In: *Social Media+ Society* 7.3 (2021), p. 20563051211042397.

[407] Pat Sable. “Anxious attachment in adulthood: Therapeutic implications”. In: *Journal of Analytic Social Work* 2.1 (1994), pp. 5–24.

[408] Pat Sable. “Attachment theory and post-traumatic stress disorder”. In: *Journal of Analytic Social Work* 2.4 (1995), pp. 89–109.

[409] Hicham Sadok, Fadi Sakka, and Mohammed El Hadi El Maknouzi. “Artificial intelligence and bank credit analysis: A review”. In: *Cogent Economics & Finance* 10.1 (2022), p. 2023262.

[410] Seref Sagiroglu and Duygu Sinanc. “Big data: A review”. In: *2013 international conference on collaboration technologies and systems (CTS)*. IEEE. 2013, pp. 42–47.

[411] Gayle Salamon. “The life and death of Latisha King”. In: *The Life and Death of Latisha King*. New York University Press, 2018.

[412] Daniel A Salmon et al. “Vaccine hesitancy: causes, consequences, and a call to action”. In: *Vaccine* 33 (2015), pp. D66–D71.

[413] Mattia Samory, Vartan Kesiz Abnousi, and Tanushree Mitra. “Characterizing the Social Media News Sphere through User Co-Sharing Practices”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14. 2020, pp. 602–613.

[414] Sara Santarossa and Sarah J Woodruff. “# SocialMedia: Exploring the relationship of social networking sites on body image, self-esteem, and eating disorders”. In: *Social Media+ Society* 3.2 (2017), p. 2056305117704407.

[415] Sapna Maheshwari. “Uncovering Instagram Bots With a New Kind of Detective Work”. In: *The New York Times* (Mar. 2018). URL: https://www.nytimes.com/2018/03/12/business/media/instagram-bots.html?searchResultPosition=1.

[416] Badrul Sarwar et al. “Item-based collaborative filtering recommendation algorithms”. In: *Proceedings of the 10th international conference on World Wide Web*. 2001, pp. 285–295.

[417] Yutaka Sasaki et al. “The truth of the f-measure. 2007”. In: *URL: https://www. cs. odu. edu/˜ mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07. pdf [accessed 2021-05-26]* (2007).

[418] Rachel Saunders et al. “Pathways to earned-security: The role of alternative support figures”. In: *Attachment & human development* 13.4 (2011), pp. 403–420.

[419] Devansh Saxena. *How algorithms are harming child welfare agencies and the Kids They Serve*. Apr. 2023. URL: https://medium.com/datasociety-points/how-algorithms-are-harming-child-welfare-agencies-and-the-kids-they-serve-596cc776c034.

[420] Devansh Saxena and Shion Guha. “Algorithmic Harms in Child Welfare: Uncertainties in Practice, Organization, and Street-level Decision-Making”. In: *ACM J. Responsib. Comput.* (Sept. 2023). Just Accepted. DOI: 10.1145/3616473. URL: https://doi.org/10.1145/3616473.

[421] Ilene Schecter. “A secure place: Attachment patterns and socioeconomic status”. In: (2013).

[422] Morgan Klaus Scheuerman, Stacy M. Branham, and Foad Hamidi. “Safe Spaces and Safe Places: Unpacking Technology-Mediated Experiences of Safety and Harm with Transgender People”. In: *Proc. ACM Hum.-Comput. Interact.* 2.CSCW (Nov. 2018). DOI: 10.1145/3274424. URL: https://doi.org/10.1145/3274424.

[423] Morgan Klaus Scheuerman et al. “A framework of severity for harmful content online”. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (2021), pp. 1–33.

[424] Milo Schield. “Information literacy, statistical literacy and data literacy”. In: *Iassist Quarterly (IQ)*. Citeseer. 2004.

[425] John R Searle. “Is the brain a digital computer?” In: *Proceedings and addresses of the American Philosophical Association*. Vol. 64. 3. JSTOR. 1990, pp. 21–37.

[426] Joseph Seering. “Reconsidering community self-moderation: the role of research in supporting community-based models for online content moderation”. In: *Proceedings of the ACM on Human-Computer Interaction* 4 (2020), p. 107.

[427] Joseph Seering et al. “Moderator engagement and community development in the age of algorithms”. In: *New Media and Society* 21 (Jan. 2019), p. 146144481882131. DOI: 10.1177/1461444818821316.

[428] Phillip R Shaver and Mario Mikulincer. “Attachment-related psychodynamics”. In: *Attachment & human development* 4.2 (2002), pp. 133–161.

[429] Renee Shelby et al. “Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction”. In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 2023, pp. 723–741.

[430] Hong Shen et al. “Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors”. In: *arXiv preprint arXiv:2105.02980* (2021).

[431] Po-Kang Shih et al. “Learning ethics in AIteaching non-engineering undergraduates through situated learning”. In: *Sustainability* 13.7 (2021), p. 3718.

[432] Donghee Shin and Yong Jin Park. “Role of fairness, accountability, and transparency in algorithmic affordance”. In: *Computers in Human Behavior* 98 (2019), pp. 277–284.

[433] Jack P Shonkoff. “From neurons to neighborhoods: old and new challenges for developmental and behavioral pediatrics”. In: *Journal of Developmental & Behavioral Pediatrics* 24.1 (2003), pp. 70–76.

[434] Eugenia Siapera. “AI Content Moderation, Racism and (de) Coloniality”. In: *International Journal of Bullying Prevention* 4.1 (2022), pp. 55–65.

[435] Mohammed Yaseen Ahmed Siddiqui et al. “Social media misinformationAn epidemic within the COVID-19 pandemic”. In: *The American journal of tropical medicine and hygiene* 103.2 (2020), p. 920.

[436] Bret L Simmons et al. “Secure attachment: Implications for hope, trust, burnout, and performance”. In: *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 30.2 (2009), pp. 233–247.

[437] Jeffry A Simpson and W Steven Rholes. “Adult attachment, stress, and romantic relationships”. In: *Current opinion in psychology* 13 (2017), pp. 19–24.

[438] Anneliese A Singh, Sarah E Meng, and Anthony W Hansen. “I am my own gender: Resilience strategies of trans youth”. In: *Journal of counseling & development* 92.2 (2014), pp. 208–218.

[439] Burrhus F Skinner. “Operant behavior.” In: *American psychologist* 18.8 (1963), p. 503.

[440] Michael E Skinner. “Promoting self-advocacy among college students with learning disabilities”. In: *Intervention in School and Clinic* 33.5 (1998), pp. 278–283.

[441] Jenny Slater. “Self-advocacy and socially just pedagogy”. In: *Disability Studies Quarterly* 32.1 (2012).

[442] Michael Slote. *The ethics of care and empathy*. Routledge, 2007.

[443] Jonathan A Smith. “Evaluating the contribution of interpretative phenomenological analysis”. In: *Health psychology review* 5.1 (2011), pp. 9–27.

[444] Lauren Smith. “Unfairness by algorithm: Distilling the harms of automated decision-making”. In: *Future of Privacy Forum*. 2017.

[445] Alison Smith-Renner et al. “No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–13. ISBN: 9781450367080. DOI: 10.1145/3313831.3376624. URL: https://doi.org/10.1145/3313831.3376624.

[446] Nathalie A Smuha. “From a race to AIto a race to AI regulation: regulatory competition for artificial intelligence”. In: *Law, Innovation and Technology* 13.1 (2021), pp. 57–84.

[447] C Riley Snorton and Jin Haritaworn. “Trans necropolitics: A transnational reflection on violence, death, and the trans of color afterlife”. In: *The Transgender Studies Reader Remix*. Routledge, 2013, pp. 305–316.

[448] Jacob Snow. *Amazons face recognition falsely matched 28 members of Congress with mugshots: ACLU*. Feb. 2023. URL: https://www.aclu.org/news/privacy-technology/amazons-face-recognition-falsely-matched-28.

[449] Cynthia Solomon et al. “History of logo”. In: *Proceedings of the ACM on Programming Languages* 4.HOPL (2020), pp. 1–66.

[450] Daniel G Solorzano and Tara J Yosso. “Critical race and LatCrit theory and method: Counter-storytelling”. In: *International journal of qualitative studies in education* 14.4 (2001), pp. 471–495.

[451] Daniel G Solórzano and Tara J Yosso. "Critical race methodology: Counter-storytelling as an analytical framework for education research". In: *Qualitative inquiry* 8.1 (2002), pp. 23–44.

[452] Gowthami Somepalli et al. "Diffusion art or digital forgery? investigating data replication in diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 6048–6058.

[453] Il-Yeol Song and Yongjun Zhu. "Big data and data science: what should we teach?" In: *Expert Systems* 33.4 (2016), pp. 364–373.

[454] Clare Southerton et al. "Restricted modes: Social media, content classification and LGBTQ sexual citizenship". In: *New Media & Society* (2020), p. 1461444820904362.

[455] Clare Southerton et al. "Restricted modes: Social media, content classification and LGBTQ sexual citizenship". In: *New Media & Society* 23.5 (2021), pp. 920–938.

[456] Wiley William Stem. "A Phenomenological Study of the Effects of Social Media Use on Minority Stress and Self-concept in LGB College Students". PhD thesis. New Mexico State University, 2020.

[457] Lynne Marie Stöven and Philipp Yorck Herzberg. "Relationship 2.0: A systematic review of associations between the use of social network sites and attachment style". In: *Journal of Social and Personal Relationships* 38.3 (2021), pp. 1103–1128.

[458] Eliza Strickland. "IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care". In: *IEEE Spectrum* 56.4 (2019), pp. 24–31.

[459] Maurice E Stucke and Ariel Ezrachi. "How digital assistants can harm our economy, privacy, and democracy". In: *Berkeley Technology Law Journal* 32.3 (2017), pp. 1239–1300.

[460] Subcommittee on Consumer Protection, Product Safety, and Data Security. *Protecting Kids Online: Instagram and Reforms for Young Users*. Washington, D.C., Dec. 2021. URL: https://www.commerce.senate.gov/2021/12/protecting-kids-online-instagram-and-reforms-for-young-users.

[461] Elisabeth Sulmont, Elizabeth Patitsas, and Jeremy R Cooperstock. "Can You Teach Me To Machine Learn?" In: *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*. 2019, pp. 948–954.

[462] Elisabeth Sulmont, Elizabeth Patitsas, and Jeremy R Cooperstock. "What is hard about teaching machine learning to non-majors? Insights from classifying instructors learning goals". In: *ACM Transactions on Computing Education (TOCE)* 19.4 (2019), pp. 1–16.

[463] Nicolas P Suzor et al. "What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation". In: *International Journal of Communication* 13 (2019), p. 18.

[464] PO Svanberg, Lisa Mennet, and Susan Spieker. "Promoting a secure attachment: A primary prevention practice model". In: *Clinical Child Psychology and Psychiatry* 15.3 (2010), pp. 363–378.

[465] Shabbir Syed-Abdul et al. "Misleading health-related information promoted through video-based social media: anorexia on YouTube". In: *Journal of medical Internet research* 15.2 (2013), e30.

[466] Anissa Tanweer et al. "Mapping for accessibility: A case study of ethics in data science for social good". In: *CoRR* abs/1710.06882 (2017). arXiv: 1710.06882. URL: http://arxiv.org/abs/1710.06882.

[467] Natalie Sui Yu Teo and Simon L Collinson. "Instagram and risk of rumination and eating disorders: An Asian perspective." In: *Psychology of Popular Media Culture* 8.4 (2019), p. 491.

[468] David W Test et al. "A conceptual framework of self-advocacy for students with disabilities". In: *Remedial and Special education* 26.1 (2005), pp. 43–54.

[469] The Instagram Team. *Community Guidelines*. URL: https://help.instagram.com/477434105621119.

[470] *the Rise and Rise of Data Science*. URL: https://www.iop.org/rise-and-rise-data-science.

[471] Melina A Throuvala et al. "Perceived challenges and online harms from social media use on a severity continuum: a qualitative psychological stakeholder perspective". In: *International journal of environmental research and public health* 18.6 (2021), p. 3227.

[472] Kim Toffoletti et al. "Visibility and vulnerability on Instagram: negotiating safety in womens online-offline fitness spaces". In: *Leisure Sciences* (2021), pp. 1–19.

[473] Suzanne Tolmeijer et al. "Female by default?–exploring the effect of voice assistant gender and pitch on trait and trust attribution". In: *Extended abstracts of the 2021 CHI conference on human factors in computing systems*. 2021, pp. 1–7.

[474] Dave Touretzky. *AI4K12*. https://github.com/touretzkyds/ai4k12/wiki. 2019.

[475] David Touretzky et al. "Envisioning AI for k-12: What should every child know about AI?" In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 9795–9799.

[476] Philipp Tschandl. "Risk of bias and error from data sets used for dermatologic artificial intelligence". In: *JAMA dermatology* 157.11 (2021), pp. 1271–1273.

[477] Sherry Turkle and Seymour Papert. "Epistemological pluralism: Styles and voices within the computer culture". In: *Signs: Journal of women in culture and society* 16.1 (1990), pp. 128–157.

[478] Jacqueline Urakami et al. "Finding Strategies Against Misinformation in Social Media: A Qualitative Study". In: *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 2022, pp. 1–7.

[479] Aleksandra Urman and Mykola Makhortykh. "Foreign beauties want to meet you: The sexualization of women in Googles organic and sponsored text search results". In: *new media & society* (2022), p. 14614448221099536.

[480] Nikki Usher, Jesse Holcomb, and Justin Littman. "Twitter makes it worse: Political journalists, gendered echo chambers, and the amplification of gender bias". In: *The international journal of press/politics* 23.3 (2018), pp. 324–344.

[481] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. "At the End of the Day Facebook Does What It Wants: How Users Experience Contesting Algorithmic Content Moderation". In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2 (2020), pp. 1–22.

[482] M-P Vaillancourt-Morel et al. "For the love of being liked: a moderated mediation model of attachment, likes-seeking behaviors, and problematic Facebook use". In: *Addiction Research & Theory* 28.5 (2020), pp. 397–405.

[483] Aimee Van Wynsberghe. "Sustainable AI: AI for sustainability and the sustainability of AI". In: *AI and Ethics* 1.3 (2021), pp. 213–218.

[484] Rebecca A VanMeter, Douglas B Grisaffe, and Lawrence B Chonko. "Of likes and pins: The effects of consumers' attachment to social media". In: *Journal of Interactive Marketing* 32 (2015), pp. 70–88.

[485] Julia Velkova and Anne Kaun. "Algorithmic resistance: media practices and the politics of repair". In: *Information, Communication & Society* 24.4 (2021), pp. 523–540.

[486] Agnes E Venema. "Deepfake Disinformation: How Digital Deception and Synthetic Media Threaten National Security". In: *Routledge Handbook of Disinformation and National Security*. Routledge, 2023, pp. 175–191.

[487] James Vincent. *The Verge Robots and AI are going to make social inequality even worse, says new report*. https://www.theverge.com/2017/7/13/15963710/robots-ai-inequality-social-mobility-study. 2017.

[488] Darshali A Vyas, Leo G Eisenstein, and David S Jones. *Hidden in plain sightreconsidering the use of race correction in clinical algorithms*. 2020.

[489] Robert J Waldinger et al. "Mapping the road from childhood trauma to adult somatization: the role of attachment". In: *Psychosomatic medicine* 68.1 (2006), pp. 129–135.

[490] Rae Walker, Jess Dillard-Wright, and Favorite Iradukunda. "Algorithmic bias in artificial intelligence is a problemAnd the root issue is power". In: *Nursing Outlook* 71.5 (2023), p. 102023.

[491] Sheridan Wall and Hilke Schellmann. *Linkedins job-matching AI was biased. The companys solution? More AI*. 2021.

[492] Yuxuan Wan et al. "Biasasker: Measuring the bias in conversational ai system". In: *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 2023, pp. 515–527.

[493] Yilun Wang and Michal Kosinski. "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images." In: *Journal of personality and social psychology* 114.2 (2018), p. 246.

[494] William Warner and Julia Hirschberg. "Detecting hate speech on the world wide web". In: *Proceedings of the second workshop on language in social media*. 2012, pp. 19–26.

[495] Anne L Washington. "How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate". In: *Colo. Tech. LJ* 17 (2018), p. 131.

[496] Thomas Way et al. "Broader and earlier access to machine learning". In: *Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education*. 2016, pp. 362–362.

[497] Thomas Way et al. "Machine learning modules for all disciplines". In: *Proceedings of the 2017 ACM Conference on Innovation and Technology in Computer Science Education*. 2017, pp. 84–85.

[498] Michael Wehmeyer, Hank Bersani, and Ray Gagne. "Riding the third wave: Self-determination and self-advocacy in the 21st century". In: *Focus on autism and other developmental disabilities* 15.2 (2000), pp. 106–115.

[499] K Wells. *An eating disorders chatbot offered dieting advice, raising fears about AI in health. 2023*.

[500] Mark West, Rebecca Kraut, and Han Ei Chew. "I'd blush if I could: closing gender divides in digital skills through education". In: (2019).

[501] *What is Artificial Intelligence (AI) ?* URL: https://www.ibm.com/topics/artificial-intelligence.

[502] *What is Data Science?* URL: https://www.ibm.com/topics/data-science.

[503] *What is Machine Learning?* URL: https://www.ibm.com/topics/machine-learning.

[504] *What's driving the demand for data scientists?* URL: https://knowledge.wharton.upenn.edu/article/whats-driving-demand-data-scientist/.

[505] Danielle Whicher et al. "Avoiding Racial Bias in Child Welfare Agencies Use of Predictive Risk Modeling". In: (2022).

[506] Norbert Wiener. "Some Moral and Technical Consequences of Automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers." In: *Science* 131.3410 (1960), pp. 1355–1358.

[507] Michelle Hoda Wilkerson and Joseph L Polman. "Situating data science: Exploring how relationships to data shape learning". In: *Journal of the Learning Sciences* 29.1 (2020), pp. 1–10.

[508] J Corey Williams et al. "Colorblind Algorithms: Racism in the Era of COVID-19". In: *Journal of the National Medical Association* 112.5 (2020), pp. 550–552.

[509] Randi Williams, Stephen P Kaputsos, and Cynthia Breazeal. "Teacher perspectives on how to train your robot: A middle school AI and ethics curriculum". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 17. 2021, pp. 15678–15686.

[510] Richard Ashby Wilson and Molly K Land. "Hate speech on social media: Content moderation in context". In: *Conn. L. Rev.* 52 (2020), p. 1029.

[511] Eleanor M Winpenny, Theresa M Marteau, and Ellen Nolte. "Exposure of children and adolescents to alcohol marketing on social media websites". In: *Alcohol and Alcoholism* 49.2 (2014), pp. 154–159.

[512] Rebecca Wong. "Guidelines to Incorporate Trauma-Informed Care Strategies in Qualitative Research". In: (2021).

[513] Yin-Ling I Wong et al. "Predicting staying in or leaving permanent supportive housing that serves homeless people with serious mental illness". In: *Departmental Papers (SPP)* (2006), p. 111.

[514] Carole-Jean Wu et al. "Sustainable ai: Environmental implications, challenges and opportunities". In: *Proceedings of Machine Learning and Systems* 4 (2022), pp. 795–813.

[515] Xingjiao Wu et al. "A survey of human-in-the-loop for machine learning". In: *Future Generation Computer Systems* 135 (2022), pp. 364–381.

[516] Sijia Xiao, Coye Cheshire, and Niloufar Salehi. "Sensemaking, Support, Safety, Retribution, Transformation: A Restorative Justice Approach to Understanding Adolescents Needs for Addressing Online Harm". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3517614. URL: https://doi.org/10.1145/3491102.3517614.

[517] Benjamin Xie, Greg L Nelson, and Amy J Ko. "An explicit strategy to scaffold novice program tracing". In: *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*. ACM. 2018, pp. 344–349.

[518] Feiyu Xu et al. "Explainable AI: A brief survey on history, research areas, approaches and challenges". In: *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*. Springer. 2019, pp. 563–574.

[519] Kyra Yee, Uthaipon Tantipongpipat, and Shubhanshu Mishra. "Image cropping on twitter: Fairness metrics, their limitations, and the importance of representation, design, and agency". In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (2021), pp. 1–24.

[520] Chi-Hsien (Eric) Yen et al. "Narratives + Diagrams: An Integrated Approach for Externalizing and Sharing People's Causal Beliefs". In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW2 (Oct. 2021). DOI: 10.1145/3479588. URL: https://doi.org/10.1145/3479588.

[521] Samuel Yeom et al. "Privacy risk in machine learning: Analyzing the connection to overfitting". In: *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. IEEE. 2018, pp. 268–282.

[522] Muhsin Yesilada and Stephan Lewandowsky. "Systematic review: YouTube recommendations and problematic content". In: *Internet policy review* 11.1 (2022).

[523] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. "Understanding the Effect of Accuracy on Trust in Machine Learning Models". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM. 2019, p. 279.

[524] Deborah R Yoder-Himes et al. "Racial, skin tone, and sex disparities in automated proctoring software". In: *Frontiers in Education*. Vol. 7. Frontiers. 2022, p. 881449.

[525] Travis Zack et al. "Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study". In: *The Lancet Digital Health* 6.1 (2024), e12–e22.

[526] Moira L Zellner et al. "Modeling, Learning, and Planning Together: An Application of Participatory Agent-based Modeling to Environmental Planning." In: *Journal of the Urban & Regional Information Systems Association* 24.1 (2012).

[527] Jing Zeng and D Bondy Valdovinos Kaye. "From content moderation to visibility moderation: A case study of platform governance on TikTok". In: *Policy & Internet* 14.1 (2022), pp. 79–95.

[528] Baobao Zhang and Allan Dafoe. "Artificial Intelligence: American Attitudes and Trends". In: *Available at SSRN 3312874* (2019).

[529] Bo Zhang, Na Wang, and Hongxia Jin. "Privacy concerns in online recommender systems: influences of control and user data input". In: *10th Symposium On Usable Privacy and Security ({SOUPS} 2014)*. 2014, pp. 159–173.

[530] Daniel Zhang et al. "The AI index 2021 annual report". In: *arXiv preprint arXiv:2103.06312* (2021).

[531] Helen Zhang et al. "Integrating ethics and career futures with technical learning to promote AI literacy for middle school students: An exploratory study". In: *International Journal of Artificial Intelligence in Education* (2022), pp. 1–35.

[532] Abigail Zimmermann-Niefield et al. "Youth learning machine learning through building models of athletic moves". In: *Proceedings of the 18th ACM International Conference on Interaction Design and Children*. 2019, pp. 121–132.

[533] John Zoshak and Kristin Dew. "Beyond Kant and Bentham: How Ethical Theories Are Being Used in Artificial Moral Agents". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445102. URL: https://doi.org/10.1145/3411764.3445102.

[534] John Zoshak and Kristin Dew. “Beyond kant and bentham: How ethical theories are being used in artificial moral agents”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–15.