

Understanding & Supporting Bottom-Up Resistance To Harmful AI Systems

Alicia DeVrio

CMU-HCII-26-105

June 2026

Thesis Committee

Kenneth Holstein (Chair), Carnegie Mellon University

Sarah Fox, Carnegie Mellon University

Jessica Hammer, Carnegie Mellon University

Richmond Wong, Georgia Institute of Technology

*Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy*

in the

Human-Computer Interaction Institute
Carnegie Mellon University

Keywords: AI resistance, AI harm, power dynamics, critical algorithm studies

This research was partially funded by the NSF, the CMU Digital Transformation and Innovation Center sponsored by PwC, the Jacobs Foundation through the Connecting the EdTech Research Ecosystem network, Apple, and the CMU-NIST AI Measurement Science and Engineering Center.

Abstract

AI systems often behave in ways that harm people, such as by perpetuating racial stereotypes, violating privacy, damaging health, and causing economic losses. There have been many efforts to address AI harms that are led top-down by researchers, industry practitioners, government officials, or other technologists. These top-down, more centrally-led approaches to AI harm remediation tend to view AI systems as inevitable and thus conceive of reformist solutions to harm.

At the same time, people affected by AI harm regularly take both small and large actions, from individual workarounds on algorithmic platforms to collectively organized protest and sabotage of algorithmic systems, to resist AI harm in the ways they can. These bottom-up forms of resistance to AI harm have broader imaginaries than top-down methods of addressing AI harm, envisioning beyond the bounds of AI system logics to see technology as part of the picture but not necessarily so.

Examining bottom-up resistance to AI harms is especially important because existing structures of power around AI systems often fail to empower those who directly interact with and are most affected by AI harm. But, there has been less work focusing on how to directly support those negatively impacted by AI systems taking action in bottom-up ways. **I argue that to make meaningful progress toward addressing the harmful impacts of AI systems, bottom-up resistance must be better understood, valued, engaged with, and broadly supported.**

In the first part of this dissertation, I examine one form of bottom-up resistance that I call *everyday algorithm auditing*, a process by which people work from the bottom up to surface, interrogate, and work toward remediation of AI harm that may elude detection via more centrally organized forms of auditing. First, I conceptualize this process to understand its dynamics and characteristics. Then, I investigate how people go about searching for and making sense of potentially harmful AI behavior from the bottom up, highlighting ways they might be further supported in their auditing activities. Finally, I examine how industry practitioners engage with bottom-up auditing, drawing out a complex relationship with fraught power dynamics between practitioners and everyday auditors.

To that end, in the bridge section of this dissertation, I then step back to examine the function and impact of algorithm audits—including bottom-up ones—in larger structures of power. This highlights a need to attend to the relationship between acts of bottom-up resistance and larger power dynamics.

Thus, in the second part of this dissertation, I take a more expansive look at how bottom-up resistance interacts with power. First, I develop a taxonomy of bottom-up resistance to AI harm that connects different acts of resistance, including everyday algorithm auditing, to existing theorizations of power. Next, I study how people understand their own power in algorithmic systems, highlighting ways to support them in more fully realizing and taking action using their power. Finally, I examine disconnects and alignments in perceptions of bottom-up resistance held by AI deployers and AI resisters, using speculative cases of algorithmic system sabotage as probes, and explore challenges and opportunities for the powerful to more productively engage with, value, and support bottom-up resistance to AI harm.

*for Meghan,
until we meet again*

Acknowledgements

I was recently reminded of a ceremony I attended at the end of high school. One by one, graduating seniors were invited to approach a microphone and announce their future plans. When it was my turn, instead of describing a college I'd committed to or a major I was excited to pursue, I said, "Hi, my name is Alicia, and I don't know!"

I've come a long way since then, learning a lot about my own interests and priorities. In graduate school alone, as part of the process of writing this dissertation, I learned how to develop my research taste, how to write for academia, how to iterate and iterate and iterate. I learned what it was like to move to a city where I knew no one to start graduate school in the midst of a pandemic. I learned what it is to lose a close friend and to gain a new family. I learned how to be lost generatively, how to wander playfully, how to embrace a little chaos. I learned through experience the importance of community support both given and received. And, of course, I learned that there is always more room for me to continue learning and growing as a researcher and as a person.

This learning would not have been possible for me without the support of many mentors, collaborators, colleagues, and friends along the way. I am immensely grateful for those who believed in me when I had little in the way of prior research experience, for those who helped me navigate the complexities of graduate school, and for those who made academia feel more welcoming and human. I aspire to share the type of support I have been lucky enough to receive.

To my advisor, **Ken** Holstein: Thank you for guidance ranging from big-picture framing all the way down to smaller word choice decisions. Your sharp eye for places of complexity has only improved my work. You helped me grow into a researcher and taught me lots about how academia works. Thank you for encouraging me to work in whatever ways worked best for me and for the freedom you provided me to develop and pursue a research agenda that deeply resonates with me.

To my committee members, Jess Hammer, Sarah Fox, & Richmond Wong: **Jess**, thank you for the directness and clarity that you bring to all your research and mentorship. Conversations with you helped me both clearly reflect on my own interests and goals and directly take action to instantiate them. I really appreciate your commitment to understanding and supporting me, especially as I wound my way around deciding whether to keep going with my PhD. **Sarah**, thank you for always approaching my work thoughtfully and thoroughly. Your feedback and sharp instinct for what's missing has repeatedly strengthened my work, and I have found the questions you ask about it to be very productive for my own thinking, pushing me in exciting directions. **Richmond**, thank you for bringing both nuance and whimsy to my research. Your methodological expertise has been invaluable, and I've really appreciated how you carefully consider my work, somehow reading it both precisely and expansively at the same time.

I'm also grateful for mentorship, advice, and support from a number of others throughout my PhD. To **Su Lin** Blodgett & **Alexandra** Olteanu: Thank you for taking me on as a Microsoft Research intern over a summer that helped me expand my understanding about what sorts of work I could do and might want to do. To **Jeff** Huang & **Peggy** Chang: Thank you for helping me embrace having a lot of seemingly very different interests during my undergraduate at Brown and advocating for me since then. To **Jamie** Boyle & **Harrison** Apple: Thank you for supporting me as a fiber artist and encouraging me to pursue opportunities related to weaving. To **Justine** Zhang: Thank you for interpreting my less than half-baked thoughts with a generous yet critical eye and for welcoming me into the UMSI political economy

reading group. To **Priyanka** Nanayakkara: Thank you for walking me through details I didn't know I didn't know in multiple long calls when I was applying for grad school, and for your friendship since then. To **Ali** Alkhatib: Thank you for helping me write and think more clearly, strongly, and critically about research and academia more broadly.

There are many who helped make the work in this dissertation what it is today. To all my collaborators on the work in this dissertation, **Wesley** Hanwen Deng, **Aditi** Dhabalia, **Motahhare** Eslami, **Sarah** Fox, **Bill** Boyuan Guo, **Jess** Hammer, **Ken** Holstein, **Hong** Shen, **Samantha** Shorey, **Cella** Sum, **David** Gray Widder, **Richmond** Wong, **Jonathan** Zong: Thank you for enriching both this work and my own research life. Without you, this work would be lesser. And to those whose experiences form the basis of this dissertation, whether as study participants or via other media sources: Thank you for sharing your stories, perspectives, thoughts, feelings, and actions. Without you, this work would not exist.

To my friends across institutions and locations, **Amina** Abdu, **Shira** Abramovich, **McKane** Andrus, **Inha** Cha, **Myra** Cheng, **Samantha** Dalal, **Jay** Foriest, **Linda** Huber, **Gena** Kim, **Nathan** Kim, **Calvin** Liang, **Jasmine** Lu, **Priyanka** Nanayakkara, **Sachita** Nishal, **Sachin** Pendse, **Renee** Shelby, **Neilly** Tan, **Joice** Tang, **Emily** Tseng, **Nick** Vincent, **Ruotong** Wang, **Jonathan** Zong: Thank you for creating warmth, safety, and a sense of community I have deeply appreciated inside academia.

To my grad school friends more nearby, **Adinawa** Adjagbodjou, **Lea** Albaugh, **Rose** Chang, **Erica** Principe Cruz, **Wesley** Deng, **Lisa** Egede, **Frank** Elavsky, **Bobbie** Fan, **Frederic** Gmeiner, **Luke** Guerdan, **Tejus** Gupta, **Noor** Hammad, **Nari** Johnson, **Shivani** Kapania, **Anna** Kawakami, **Pranav** Khadpe, **Yasmine** Kotturi, **Tzu-Sheng** Kuo, **Franklin** Li, **Sanika** Moharana, **Katelyn** Morrison, **Tucker** Rae-Grant, **Vivian** Shen, **Venkat** Sivaraman, **Jaemarie** Solyst, **Samantha** Speer, **Cella** Sum, **Alice** Tang, **Jordan** Taylor, **Kimi** Wenzel, and **David** Widder: thank you for making Pittsburgh feel like home. Thanks especially to **Franky** Spektor for many long, lazy conversations that made me feel truly understood, both academically and personally.

To the weaving friends and supporters I have met along the way, **Cathryn** Amidei, **Harrison** Apple, **Jamie** Boyle, **Luca** Budofsky, **Miranda** Crotsley, **Margaret** Dugger, **Rebecca** Hebert, **Kathie** Hollingshead, **Sarah** Khadraoui, and **Sylvia** Rhor: Thank you for inviting me into the far-reaching realms of fiber arts, for encouraging me to look closer at the textiles that surround me every day, and for modeling an enthusiasm for making detached from productivity. To my runner friends at Hounds, **Jen** Bigham, **Nicole** Feng, **Laura** Harnish, **Gretchen** Swabe, and **Andrea** Imhof: Thank you for the shared miles and smiles.

I am also extremely grateful for my far-away friends and family. To **Sajant** Anand, **Adam** Becker, **Peter** Chang, **Luci** Cooke, **Rebecca** DeVos, **Jason** Goettisheim, **Sarah** Johnson, **Sarita** Lee, **Sonia** Lee, and **Divya** Mahadevan: Thank you for the many miles travelled and couches/beds/air mattresses offered over the years; I'm glad we're never farther than a phone call away. To **my extended family** spread across this country and the world: Thank you for celebrating holidays, for hosting and visiting, for grounding me, and for sharing in my journey to and through grad school, even if I didn't always explain it the most clearly. To **my parents**: When I was little, people often asked me if I wanted to be a doctor when I grew up, and I always said, "No!"—I suppose I changed my mind; thank you for not rubbing it in. And finally, to **Nathan**: Thank you for meeting me on the corner of Fifth & Negley for a walk and, later on, for smashing your last name together with mine—grad school was a doozy, but for you, I'd do it again.

Contents

Abstract	ii
Acknowledgements	iv
1 Introduction	1
1.1 Overview & Contributions	2
1.1.1 Part I: Everyday Algorithm Auditing	2
1.1.2 Bridge: Audits & Power	3
1.1.3 Part II: Power Dynamics & Bottom-Up Resistance	3
1.1.4 Overall Contributions	4
1.2 Interpreting the Contributions	4
1.2.1 Positionality	4
1.2.2 Key Concepts	5
1.3 Prior Publications	6
2 Related Work	7
2.1 Algorithm Auditing To Address Algorithmic Harm	7
2.1.1 Existing Approaches to Algorithm Auditing	7
2.1.2 Limitations of Existing Auditing Approaches	8
2.2 The Importance of Power Dynamics for Addressing Algorithmic Harm	9
I Everyday Algorithm Auditing	11
3 Understanding the Power of Everyday Users	12
3.1 Introduction	12
3.2 Everyday Algorithm Auditing	15
3.2.1 Scope and Boundaries	15
3.2.2 Everyday Algorithmic Resistance	16
3.2.3 Counterpublics	17
3.3 Methods	18
3.4 Case Studies	20
3.4.1 Twitter Image Cropping Algorithm	20
3.4.2 Rating Platform Algorithms	23
3.5 The Lifetime and Dynamics of an Everyday Audit	26
3.5.1 Initiation	26
3.5.2 Awareness Raising	26
3.5.3 Hypothesizing & Testing	27
3.5.4 Remediation	28
3.6 Design Implications	29
3.6.1 Design for Community Guidance	30
3.6.2 Design for Expert Guidance	30
3.6.3 Design for Algorithmic Guidance	31

3.6.4	Design for Organizational Guidance	32
3.6.5	Design for Incentivization	32
3.7	Discussion	33
3.7.1	The Power of Everyday Algorithm Auditing	34
3.7.2	Appropriate and Timely Interventions	35
3.8	Conclusions	35
4	User Strategies for Uncovering Algorithmic Harm	37
4.1	Introduction	37
4.2	Related Work	39
4.2.1	User-Driven Algorithmic Auditing	39
4.2.2	Information Search & Sensemaking	40
4.3	Methods	40
4.3.1	Recruitment	41
4.3.2	Participants	41
4.3.3	Bias & Harm	42
4.3.4	Phase 1 – Think-Aloud Interview: Individual bias search & sensemaking	42
4.3.5	Phase 2 – Diary Study: Finding harmful algorithmic behavior in everyday life	44
4.3.6	Phase 3 – Workshops: Collective evaluation & validation of reported biases	44
4.4	Results	46
4.4.1	Users’ Power in Detecting Anticipated & Potential Biases & Harms	46
4.4.2	Developing a Process Model of Bias Search & Sensemaking	48
4.4.3	Influences on Bias Search & Sensemaking	48
4.4.4	Stage 1: Search Inspiration	51
4.4.5	Stage 2: Sensemaking	52
4.4.6	Stage 3: Remediation	57
4.5	Limitations	60
4.6	Discussion	60
4.6.1	Supporting User-Driven Algorithm Auditing	60
4.6.2	Limits & Challenges of User-Driven Algorithm Auditing	63
4.6.3	Algorithmic Harm Remediation: What’s next?	64
4.7	Conclusion	65
5	Industry & User-Engaged Algorithm Auditing	66
5.1	Introduction	66
5.2	Related Work	68
5.2.1	Understanding and supporting responsible AI practices in in- dustry contexts	68
5.2.2	The power of users in algorithm auditing	69
5.2.3	Supporting user-engaged algorithm auditing	70
5.3	Method	71
5.3.1	Study design	71
5.3.2	Participants	73
5.3.3	Data analysis	74
5.4	Findings	75
5.4.1	Practitioners’ motivations and practices for engaging users to audit AI products and services	75

5.4.2	Challenges and needs in user-engaged algorithm auditing . . .	79
5.4.3	Organizational obstacles to user-engaged algorithm auditing	83
5.5	Limitations	85
5.6	Discussion	85
5.6.1	Unique challenges in supporting user-engaged algorithm auditing	85
5.6.2	The complex relationship between industry practitioners and user auditors	87
5.6.3	Users' perspectives on user-engaged algorithm auditing	88
5.7	Conclusion	89
Bridge: Audits & Power		90
6	Power-Related Pitfalls of Algorithm Audits	91
6.1	Introduction	91
6.2	What are algorithm audits?	94
6.3	Pitfalls of algorithm audits	95
6.3.1	Algorithm audits further the concentration of power.	95
6.3.2	Algorithm audits limit how both problems and solutions are envisioned.	97
6.3.3	Algorithm audits preserve a facade of positive change.	98
6.4	Ways Forward: Grappling with power toward a more effective theory of change	99
6.4.1	Embracing new commitments	100
6.4.2	Scrutinizing the accountability ecosystem	101
6.5	Conclusion	103
II Power Dynamics around Bottom-Up Resistance		104
7	A Taxonomy of Responses From Below to Algorithmic Harm	105
7.1	Introduction	105
7.2	Related Work	107
7.2.1	Responses From Below to Algorithmic Harm	107
7.3	Taxonomy of Responses from Below to Algorithmic Harm	107
7.3.1	Process of Response Collection & Categorization	107
7.3.2	Mitigating Individual Algorithmic Harms	109
7.3.3	Pursuing Legal Avenues Against Algorithmic Harm	110
7.3.4	Investigating Potentially Harmful Algorithmic Behaviors Further	111
7.3.5	Refusing Engagement With Harmful Algorithmic Systems . . .	111
7.3.6	Communicating Algorithmic Harm With Others	114
7.4	Limitations	116
7.5	Discussion	117
7.5.1	Burdens of Responding to Algorithmic Harms From Below . .	117
7.5.2	(Adverse) Impacts of Responses From Below to Algorithmic Harm	118
7.5.3	Obstacles Faced When Responding From Below to Algorithmic Harms	119
7.6	Conclusion	120

8	Algorithmic Users' Conceptions of Power	121
8.1	Introduction	121
8.2	Related Work	122
8.2.1	Power dynamics & games	122
8.2.2	Users' power & games	123
8.3	Methods	124
8.4	Results	126
8.4.1	Users perceive their power as limited, individualistic: The need for systems thinking	126
8.4.2	Sharing support facilitates connection and control: The need for collective action	128
8.4.3	Feelings of safety enhance agency: The need for play & explo- ration	129
8.5	Limitations	131
8.6	Discussion	132
8.6.1	Supporting systems thinking through games	132
8.6.2	Supporting collective action through games	134
8.6.3	Supporting play & exploration through games	135
8.7	Conclusion	137
9	Saboteur & Deployer Perceptions of Algorithmic Sabotage	138
9.1	Introduction	138
9.2	Related Work	141
9.2.1	Breaking AI systems & testing	141
9.2.2	AI refusal & resistance	141
9.3	Methods	142
9.4	Results	146
9.4.1	Sabotage as feedback to improve algorithmic systems?	146
9.4.2	Sabotage as design?	148
9.4.3	Sabotage aims to propel resistance movements against prob- lematic algorithmic system logics, but deployer responses to sabotage frame the system logics as inevitable	152
9.5	Limitations	156
9.6	Discussion	156
9.6.1	Supporting AI sabotage as communicative practice	157
9.6.2	Supporting AI sabotage as design practice	158
9.6.3	Supporting AI sabotage as ideological practice	159
9.7	Conclusion	160
10	Conclusion	161
A	Ch. 4 Demographic information	163
B	Ch. 7 Additional Details on Process of Response Collection & Categoriza- tion	164
	Bibliography	166

List of Figures

3.1	High-level descriptions of our four primary case studies across the three dimensions outlined in the <i>Scope and Boundaries</i> section. These cases serve as illustrative examples to better understand how everyday audits can work, what paths they can take, and what impacts they can have. Darker shades represent higher levels along a given dimension; lighter shades represent lower. Note that less variation is presented along the “Organicness” dimension given that this chapter focuses on everyday audits toward the more organic end of the spectrum as less organic everyday audits have been well described by past literature.	21
4.1	One example case submitted by our diary study participant P3 (“romantic couples”), including a screenshot of the potential bias, how this case was found, why this case might be harmful and whom it might harm.	47
4.2	Process model depicting stages of and influences on users’ bias search and sensemaking process. Bold boxes and text represent the three major stages of the process; other bubbles and text represent influences on and impacts of this process. Larger arrows represent movement between stages of the process, while the smaller arrows represent interactions between the process stages and influences on the process (i.e., platform affordances, user knowledge and beliefs).	48
5.1	Two potential user-engaged auditing pipeline designs that were iteratively co-designed with participants. The left image shows the “developer-led” pipeline design, and the right image shows the “user-led” pipeline design. Each figure illustrates a possible interaction flow between user auditors and AI product teams, showing how auditing tasks are created, how background information on user auditors is shared, how user auditing reports are generated based on auditors’ findings, and how these reports are shared with AI product teams. During the co-design activity, participants could zoom in, annotate, and modify the details. We used these pipeline flowcharts as probes, not as final products , to investigate more deeply on practitioners’ challenges and desires.	71
5.2	High-level overview of our findings. We first share participants’ existing motivations and practices for user engagement around algorithm testing and auditing (Section 5.4.1). We then describe challenges participants have faced in their attempts to effectively engage users (Section 5.4.2). Finally, we share broader organizational obstacles to user-engagement around algorithm auditing perceived by participants, highlighting key tensions that arise in practice when involving users in algorithm auditing efforts (Section 5.4.3).	75

A.1 Self-reported participant demographic information from the screening questionnaire.	163
A.2 Participant familiarities with various algorithmic concepts, on a 5-point Likert scale.	163

List of Tables

3.1	15 cases of everyday algorithm auditing across various domains, which we use to ground our discussion throughout this paper. Bold cases are analyzed in depth.	19
4.1	Detailed descriptions of the four primary cases (user reports from our diary study) selected by workshop participants.	46
5.1	Summary of participants' backgrounds and relevant experience.	73
5.2	Participants' existing practices for testing and auditing their AI systems for harmful behaviors and biases. In this table, we report both approaches that do not engage users and user-engaged approaches. Section 5.4.1 describes the <i>limitations</i> of approaches that do not engage users. Section 5.4.1 describes the user-engaged approaches in detail. In this table, approaches that do not engage users are color-coded in gray; approaches for user-engagement at small scales are shown in light green; and approaches for user-engagement at larger scales are shown in darker green.	77
7.1	Types of responses (bold), their strategies that interact with power (X), and examples	108

Chapter 1

Introduction

With the increasing integration of algorithmic systems into everyday life has also come the production, perpetuation, and amplification of harms and other negative impacts on people who use and interact with these algorithmic systems [473]. For example, Google’s algorithmic search engines have displayed sexism and racism in their search results [379]; police use of facial recognition has led to wrongful arrests [341]; Amazon has algorithmically recommended groups of merchandise expedient for killing oneself that have been called “suicide kits” [250]; and companies like Waymo have put self-driving cars on the road that block busses and emergency vehicles [338], create more traffic [198], and increase surveillance [365].

Significant efforts in academia and industry have aimed to address AI harms. For example, efforts to detect biased, discriminatory, or otherwise harmful behaviors in algorithmic systems are known as algorithm audits, and are typically led top-down by researchers, industry practitioners, government officials, or other technologists. Trying to address concerns about their vehicles, Waymo had an audit conducted by an independent technical inspection service company [515]. And to address issues of blocking emergency vehicles, Waymo developed a training for first responders [269].

These top-down, more centrally-led approaches to AI harm remediation tend to view AI systems as inevitable and thus conceive of reformist solutions to harm that do not substantially alter the logics of the AI system. For example, Waymo’s audit excluded “remote driving, dispatching, and customer support to occupants or other road users”, meaning that it completely failed to consider incidents when remote employees need to step in to, for instance, guide the autonomous vehicles past emergency vehicles. Similarly, Waymo’s safety training aims to change how first responders behave, rather than make any adjustments to the design or deployment of their self-driving cars.

At the same time, people affected by AI harm have been taking action themselves from the bottom up, in ways large and small, to resist AI harm in the ways they can. For example, people have been protesting, signing petitions, and attending public meetings to voice their opposition to self-driving car expansion [24]—and they also discovered that putting cones on self-driving car hoods makes them inoperable until an employee comes in person to address it [396], so they’ve been doing that, too. To name some other examples, TikTok users who had their posts unjustly restricted by moderation algorithms invented new forms of language to work around the harmful system [498]. High-school students who had not consented to be monitored by facial recognition unplugged the behavior-monitoring cameras before their exams, stopping the surveillance [592]. And Twitter users who extensively raised awareness about racial bias in the platform’s image cropping algorithm caused Twitter to retire the algorithm [85].

These forms of bottom-up resistance work to create material change in the world, allowing people to begin to move toward the worlds they want using what power they have. Additionally, bottom-up resistance has broader imaginaries than top-down methods of addressing AI harm; these acts envision beyond the bounds of AI logics, seeing technology as part of the picture but not necessarily so. This means that bottom-up resistance can push in directions that more fundamentally change how AI is deployed—including changing *whether* it is deployed.

Examining bottom-up resistance to AI harms is especially important because existing structures of power around AI systems often fail to empower those who directly interact with and are most affected by AI harm [22, 284, 364]. I argue that we need to center these acts of bottom-up resistance if we truly care about mitigating AI harm—that **to make meaningful progress toward addressing the harmful impacts of AI systems, bottom-up resistance must be better understood, valued, engaged with, and broadly supported.**

1.1 Overview & Contributions

This dissertation provides enhanced understanding of how people engage in forms of bottom-up resistance to AI. Bottom-up resistance can effect change, but it implicitly puts the burdens for change on affected people themselves, when really the burdens should be on those with power over AI systems. To that end, my work also provides directions moving forward for how researchers and technologists might better support bottom-up resistance.

1.1.1 Part I: Everyday Algorithm Auditing

In **Part I** of this dissertation, I present my work on one form of bottom-up resistance, which I call *everyday algorithm auditing*.

In Chapter 3, I propose and explore the concept of everyday algorithm auditing, a process in which users detect, understand, and interrogate problematic machine behaviors *via their day-to-day interactions* with algorithmic systems. To do this, I analyze real-world cases of everyday algorithm audits and argue that users are powerful in surfacing problematic machine behaviors that may elude detection in more formal forms of audits. This work contributes a conceptualization of how bottom-up resistance can have unique, often effective abilities to respond to harm emanating from algorithmic systems; an argument for expanding the scope of what research communities might understand as legitimate and useful forms of auditing; and design directions for future work to support everyday algorithm auditing.

In Chapter 4, I investigate how people go about searching for and making sense of harmful algorithmic behavior, both individually and collectively, through a series of think-aloud interviews, diary studies, and group workshops. The study's findings indicate that users' search strategies and interpretations are heavily guided by their personal experiences with and exposures to societal bias, and that collective sense-making amongst multiple users is invaluable in user-driven algorithm audits. This work contributes the first in-depth empirical investigation of how users find and make sense of harmful behaviors in algorithmic systems, both individually and collectively; a conceptual process model capturing the ways that users find and make sense of biased and harmful algorithmic behaviors, developed through an analysis of the study participants' actions and explanations; and directions for the design

of future methods, platforms, and tools to support user-driven algorithm auditing, informed by the process model.

In Chapter 5, I examine how industry practitioners view and employ user-engaged auditing practices themselves through a series of interviews and iterative co-design activities. This work contributes an in-depth understanding of industry practitioners' motivations, current practices, and challenges in their engagement with user-engaged auditing. Additionally, this work contributes insights into the complex relationship between practitioners and user auditors and suggests directions and design opportunities based on these insights for future work.

1.1.2 Bridge: Audits & Power

In the Bridge, I take a broader view of algorithm audits, considering their function and impact within larger power structures.

In Chapter 6, I argue that algorithm audits are fundamentally limited mechanisms to address the harmful impacts of algorithmic systems, as they give the appearance of addressing algorithmic harm without meaningfully challenging the systems they claim to scrutinize. To do this, I examine and bring attention to three interrelated pitfalls of algorithm auditing: that algorithm audits further the concentration of power, that algorithm audits limit how both problems and solutions are envisioned, and that algorithm audits preserve a facade of positive change. This work contributes an argument that work on AI accountability should shift focus from algorithm audits to the broader systems of power in which audits and other accountability sit, a direction I take on myself in Part II.

1.1.3 Part II: Power Dynamics & Bottom-Up Resistance

In Part II of this dissertation, I focus on better understanding the dynamics of power surrounding bottom-up resistance to harmful AI systems. I first work to explore how acts of bottom-up resistance interact with power, then investigate how resisters themselves understand their power, and close with a comparison between technologist and resistor views of bottom-up resistance.

In Chapter 7, I explore acts bottom-up resistance to AI harm by collecting documented, historical cases and developing a taxonomy. The taxonomy connects different types of resistance tactics to existing theorizations of power and centers how people employ, shift, and build power in bottom-up resistance by protecting themselves, pressuring those in control of algorithmic systems, and strengthening their capacity by building their reserves of power. This work contributes a broader, overarching understanding of how structurally disempowered people resist algorithmic harm in ways that interact with power; an argument that researchers should actively work with affected people to further their efforts to defend themselves from algorithmic harm; and opportunities for future work to support efforts from structurally disempowered people to resist algorithmic harm.

In Chapter 8, I investigate how users understand their own power to influence algorithmic systems through interviews with users of algorithmic social media platforms. The findings indicate that users perceive their power as limited and individualistic, that sharing support facilitates user connection and control, and that feelings of safety can enhance user agency. Drawing upon research foregrounding how games can support understanding of power dynamics in complex systems, this work contributes not only increased understanding of how people conceptualize their own powers in algorithmic systems but also opportunities and directions for

the design of future games to help algorithmic system users with systems thinking, collective action, and exploration and play that can support users in more fully realizing and taking acts of bottom-up resistance using their power.

In Chapter 9, I focus on cases of algorithmic sabotage as a form of bottom-up resistance that is often dismissed by those with power over algorithmic systems as immoral, unreasonable, and unhelpful. I investigate how saboteurs themselves compared to deployers view acts of algorithmic sabotage. The findings indicate that deployers understand sabotage as feedback to improve their systems and often describe saboteurs as uninformed and unimaginative, while, in contrast, saboteurs described sabotage as a way to enact alternatives to systems, rejecting engagement with deployers they viewed as unreceptive to feedback that challenged their agenda. This work contributes an in-depth empirical investigation of how saboteurs and deployers view algorithmic sabotage; a rich, nuanced understanding of disconnects and alignments between their perceptions of algorithmic sabotage and its aims; and, informed by the results, implications for researchers and technologists might better support algorithmic sabotage and resistance more broadly as communicative, design, and ideological practice.

1.1.4 Overall Contributions

In summary, this dissertation contributes the following:

- Conceptualization of how bottom-up resistance provides unique, often effective abilities to respond to harmful AI systems
- Empirical knowledge of the ways that people engage in bottom-up resistance to harmful AI systems
- Increased understanding of how people understand their own powers in algorithmic systems
- Increased understanding of how those with power over algorithmic systems view and engage with bottom-up resistance to algorithmic harm
- Design opportunities for how HCI researchers and other technologists might better prioritize and support actions of bottom-up resistance to algorithmic harm

1.2 Interpreting the Contributions

In this section I provide necessary context for how to interpret and understand this dissertation.

1.2.1 Positionality

Here I provide a reflexive positionality statement, in line with suggestions from [327]. In this dissertation, I attempt to understand existing efforts of bottom-up resistance to harmful AI systems. As an academic, I have seen and experienced the ways in which power shapes discourse, values, and priorities; as such, I focus here on the ways in which bottom-up resistance to algorithmic harm can elucidate how power dynamics are experienced and navigated by their actors, as well as the ways in which

these responses might be further supported. As a junior scholar, I wield comparatively less power within academia while holding relatively high power as a technologist and researcher to influence the ways decisions around algorithmic systems are made and considered. I understand my position as imbued with responsibility to use my work to shift power to others who are directly affected by algorithmic systems—a group that includes me, as I too have experienced some of these systems’ harms firsthand—yet have little control over these systems. I aim to do this by bringing together and strengthening existing currents of thought and action around bottom-up resistance to AI systems. At the same time I acknowledge the limitations of research as a vector for change; as such, I understand my research to be speaking primarily to other researchers and technologists. Additionally, my aspirations to represent complexity in the dynamics of and around bottom-up resistance to algorithmic harm are limited by my own set of experiences: my training as a computer scientist, writer, and human-computer interaction specialist in the United States has been conditioned by the implicit and explicit normalization of Western culture and hegemony where I currently live and work, even as I resist it. And my current work home at Carnegie Mellon University has both afforded me additional legitimacy and thus access in technological research and practitioner spaces while also surely shaped my approach in more computer science-centric ways. As such, I look forward to the ways in which others might expand, critique, reshape, and build on this work in ways that shift power even more to those who take bottom-up action to resist algorithmic harm.

1.2.2 Key Concepts

What follows is brief additional context about the understandings of key terms frequently used in this work.

First, *algorithmic systems* are understood as ideological projects, for which the impacts and uses are as if not more important than the technical details. This follows Ali Alkhatib’s contention that “AI is a political artifact through and through” [20] and, similarly, Eryk Salvaggio’s contention that “AI as an ideology is less about *what the tech does* and more about *where the tech is doing it*” [445]. Additionally, based on the lack of clear, universal definitions of algorithms and AI as well as the frequent conflation of the terms what they represent, I use “algorithms” and “AI” interchangeably throughout this dissertation [339, 503].

Next, I understand *algorithmic harm* as wrongful and harmful algorithmic system impacts [152], for which those affected are the arbiters. That is, I take an expansive view of harm, referring to those harmed to define for themselves what counts as harm. Further, harm is conceptualized in relation to power in this work and as such connects to ideas of injustice, structural violence, and domination [95, 340]. This follows Jenna Burrell’s notions of algorithmic injustice as domination, of “the denial of self-determination and control” [95].

I understand *bottom-up resistance to AI* as practices that structurally disempowered people engage to oppose the harmful impacts of AI systems and create more just sociotechnical spaces. I draw on Michel de Certeau’s and James C Scott’s ideas of “everyday resistance” to examine the ways that relatively powerless people exert power and agency in the face of dominant ruling structures in their everyday lives [139, 460]. I bring together related HCI conceptualizations that are focused on technology, specifically “routine infrastructuring” in which people build everyday resilience around algorithmic systems [466] and “tactics from below” in which people build power over technology when that power has not been given to them [568].

Throughout this dissertation I understand *power* as a contextually-based, multifaceted concept, largely informed by feminist standpoint theory [242], in which power can emanate from multiple sources and is fundamentally about structurally influenced relationships between different groups of people [305]. HCI researchers have highlighted the importance of social and political theory conceptualizations of power as “power-over others” and as “power-to do” [23, 456]. I adopt these framings, focusing on the ways that everyday people retain, maintain, and create both 1) ways to get others to engage in actions that diminish algorithmic harms that they encounter and 2) ways to act directly to shape their experiences with algorithmic systems. Following past work on power within and around algorithmic systems [284, 364], I focus on the structurally influenced relationships between different groups of people [23]. This power-aware view enables deeper understanding of how everyday people, responding to algorithmic harm, seek to interact with, garner, and use power. Notably, when advocating for shifting power or empowering, I mean this as a contention to push for systemic changes: following [121] the goal is not to assimilate everyday people affected by algorithmic harm into dominant structures but rather to reshape those dominant structures—and understand how they are being reshaped already by everyday people.

1.3 Prior Publications

The work in this dissertation not only involved but required the time, energy, insights, and care of many others. Throughout, regardless of whether I use singular or plural first-person pronouns, the work presented should be understood as a collective effort produced with the support of collaborators without whom the work would not have been the same, or simply would not have been.

To that end, much of the work that this dissertation includes and draws from has been published or is in submission or preparation as follows:

- Hong Shen*, Alicia DeVos*, Motahhare Eslami, Kenneth Holstein. *Everyday Algorithm Auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors*. CSCW, 2021.
- Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, Motahhare Eslami. *Toward User-Driven Algorithm Auditing: Investigating users’ strategies for uncovering harmful algorithmic behavior*. CHI, 2022.
- Wesley Hanwen Deng, Bill Boyuan Guo, Alicia DeVrio, Hong Shen, Motahhare Eslami, Kenneth Holstein. *Understanding Practices, Challenges, and Opportunities for User-Engaged Auditing in Industry Practice*. CHI, 2023.
- Alicia DeVrio, David Gray Widder. *No More Plaudits for Algorithm Audits: A Provocation*. AIES, 2026.
- Alicia DeVrio, Motahhare Eslami, Kenneth Holstein. *Building, Shifting, & Employing Power: A taxonomy of responses from below to algorithmic harm*. FAccT, 2024.
- Alicia DeVrio, Jessica Hammer, Kenneth Holstein. “Wow, we can’t do shit”: Opportunities to support social media users’ conceptions of power with games. In Prep.
- Alicia DeVrio, Cella Sum, Samantha Shorey, Jonathan Zong, Richmond Wong, Sarah Fox, Kenneth Holstein. *Algorithmic Sabotage: Understanding disconnects and alignments between saboteur & employer perceptions*. In Prep.

Chapter 2

Related Work

Algorithmic harm is "the adverse lived experiences resulting from [an algorithmic] system's deployment and operation in the world" [473]. With the immense amount of harm produced by algorithmic systems has come considerable work aiming to understand these harms, teasing out the distinctions around types of algorithmic harm in order to direct algorithmic harm mitigation efforts in more fruitful directions (e.g., [25, 74, 108, 453, 473, 565]). For instance, one recent taxonomy outlines five high-level categories of algorithmic harm—representational, allocative, quality-of-service, interpersonal, and social system harms—with 20 subcategories [473]. In this chapter I overview first some existing work related to addressing algorithmic harm via auditing, then some work on the importance of centering power relations when addressing algorithmic harm.

2.1 Algorithm Auditing To Address Algorithmic Harm

Many efforts—in academia, industry, governmental sectors, and beyond—have attempted to ensure that algorithmic systems adhere to notions of fairness in order to root out algorithmic harms such as bias (e.g., [6, 126]). For example, a well-known distinction between allocative and representational harms arose from research on algorithmic bias [50, 129].

A growing body of work in human-computer interaction, machine learning, artificial intelligence, and other related communities has proposed tools, processes, and frameworks to audit algorithmic systems for biased, discriminatory, or otherwise harmful behaviors (e.g., [127, 323, 417, 447]). However, these methods often fail to surface serious issues that everyday users of algorithmic systems are quickly able to detect once a system is deployed in the wild, which I explore more in Part I. In this section, I briefly overview existing approaches to algorithm auditing, an area that has received increasing attention from the CSCW, HCI, and ML communities in recent years.

2.1.1 Existing Approaches to Algorithm Auditing

Scholars have proposed a number of approaches to audit algorithmic systems for biased, discriminatory, or otherwise harmful behaviors, often under the broad umbrella of "auditing algorithms" [447]. Drawing from a wide range of methods in areas such as investigative journalism, information security and finance, and housing and employment audits, this line of work often involves third-party external experts (e.g., researchers, technologists, or policymakers) in assessing the alignment of deployed algorithmic systems with laws and regulations, societal values, ethical desiderata, or industry standards (e.g., [91, 151, 414, 447]). Meanwhile, a growing body of work has proposed tools, processes, and frameworks for internal algorithm

audits, conducted by ML teams themselves, with the aim of detecting and mitigating misalignments prior to deployment (e.g., [127, 256, 323, 417]).

Past research in this domain has uncovered harmful biases across a wide range of algorithmic systems such as search engines [312, 379, 431], housing websites [40], hiring systems [110, 136], and e-commerce platforms [241] using different kinds of approaches led by experts and researchers. For example, Robertson et al. used a survey and web browser extension to collect users' experiences in order to audit partisan audience biases in Google search results [431]. In another example, Hannak et al. used web scraping techniques and hired Amazon MTurk users as testers in order to audit a variety of top e-commerce sites for price steering and discrimination [241]. These are just a few examples of many auditing efforts that have been conducted during recent years on algorithmic systems.

Inspired by early methods of the social scientific "audit study" [366], Sandvig et al. [447] proposed a taxonomy to summarize different algorithm auditing methods and research designs, including 1) code audits, 2) noninvasive user audits, 3) scraping audits, 4) sock puppet audits, and 5) crowdsourced/collaborative audits. A *code audit* typically requires auditors to have secure access to the source code and system design. A *noninvasive user audit* might use methods like surveys to collect and synthesize users' experiences in order to support inferences about the underlying operations of an algorithmic system. A *scraping audit* involves researchers sending out repeated queries to test how an algorithmic system behaves under a variety of conditions. In a *sock puppet audit*, researchers generate fake user accounts, or "sock puppets," to investigate how an algorithmic system may behave differently in response to different user characteristics or patterns of behavior. Finally, in a *crowdsourced/collaborative audit*, instead of using "sock puppet" accounts, researchers might hire crowdworkers as testers via crowdsourcing platforms like Amazon MTurk to work on decomposed subtasks.

2.1.2 Limitations of Existing Auditing Approaches

Past literature has generally focused on expert-led algorithm auditing approaches that are performed outside the context of everyday use of an algorithmic system. Conducted as research projects or professional services, most algorithm auditing approaches that have been discussed in the literature require individuals with some level of technical expertise to initiate, conduct, or direct the entire process. For example, even approaches that rely on crowdsourcing techniques involve some centralized organization. In such crowdsourced/collaborative audits, it is still the researchers' responsibility to design and initiate the audit by assigning crowdworkers specific tasks, and then to convert the crowd's outputs on distributed tasks into meaningful insights about an algorithmic system's behavior.

However, such centrally-organized, formal audits often fail to surface serious issues that everyday users of algorithmic systems are quickly able to detect once a system is deployed in the wild. Many harmful algorithmic behaviors are challenging to anticipate or detect outside of authentic, situated contexts of use, for a variety of reasons. For example, certain algorithmic behaviors may only arise, or may only be recognized as harmful, when a system is used in the presence of particular real-world social or cultural dynamics [256, 343, 462, 465]—dynamics that may be difficult or infeasible to predict and simulate in an artificial setting.

Other harmful behaviors may only emerge when a system is used in unanticipated ways or in unanticipated contexts, perhaps due to changing norms and practices surrounding the use of a given algorithmic system over time [127, 190, 465].

One highly publicized example is Microsoft’s AI chatbot Tay, which was set up to learn over time based on its interactions with users on Twitter. In less than 24 hours of exposure to Twitter users, who began interacting with Tay in ways its developers had not foreseen, the bot began exhibiting misogynistic and racist behaviors and ultimately needed to be shut down [549]. These examples align with the broad notion of “emergent bias,” discussed in Friedman and Nissenbaum’s seminal work on bias in computer systems [190].

Existing auditing approaches can also fail to detect serious issues when those conducting an audit lack the relevant cultural background and lived experience to recognize or know where to look for harmful behaviors [589]. For example, Holstein et al. [256] found that even when ML product teams thoroughly audit their systems using existing approaches, they often are subsequently blindsided by user complaints about myriad issues that their investigations had not revealed. In one case, even after devoting many months to identifying and mitigating gender and racial biases in their image captioning system, a product team learned about a wide array of additional issues that their efforts had ignored (e.g., mosques, synagogues, and other religious sites being labelled as “churches”) only after the system was deployed and in use across a range of real-world contexts. Although crowdsourced/-collaborative auditing approaches invite participation from a larger number of people, such audits typically rely on crowdworkers, who do not necessarily represent the demographics of a given algorithmic system’s user base, and who are participating in structured tasks outside of their regular interactions with technology. Without engaging diverse, contextually situated users in the process of auditing complex algorithmic systems, existing approaches are likely to suffer major blindspots, with critical issues surfacing only post-deployment (e.g., [127, 169, 190, 256, 465]).

2.2 The Importance of Power Dynamics for Addressing Algorithmic Harm

Researchers have highlighted that we cannot productively address harmful algorithmic behaviors like bias without centering systems of power [22, 69, 74, 284, 299, 364]. For example, Miceli et al. argue that issues of power have not yet been sufficiently centered and explain how seeing the world through only the lenses of bias and fairness neglects power [364]. Alkhatib similarly contends that algorithmic harm derives from structural power and thus harm reduction requires withdrawing power from algorithmic systems [22]. In this dissertation, I focus on power as a way to get at issues of algorithmic harm.

This focus is especially important since existing structures around algorithmic systems often fail to empower those who directly interact with and are affected by algorithmic harm [22]. In response to such power imbalance, people affected by algorithmic harm frequently adopt workarounds and counter-practices to algorithmic systems in place. Recent research and real-world case studies indicate that responses “from below”—i.e., from members of the public or others without direct power to make decisions about the design or deployment of algorithmic systems [568]—can be effective in garnering power and working toward harm remediation (see Part I, [313, 552, 568]).

Organizations in charge of algorithmic systems often fail to address algorithmic harms sufficiently of their own impetus [250, 414, 530, 562]. Frequently companies respond to concerns surrounding algorithmic harm only when faced with significant evidence and negative repercussions—and even then, often only provide empty

statements. For example, families who criticized Amazon’s auto-recommendations of materials used for suicide were acknowledged with an Amazon statement that “it made customer safety a top concern” [250]. However, Amazon continued to sell the relevant products, claiming they were not responsible for customers’ misuse [530]. Clearly those affected cannot always count on organizations to remediate harms they introduce with their algorithmic systems. Structural power often fails to empower those who directly interact with and are affected by algorithmic systems, rarely allowing them input into these systems [22]. Even when everyday people are enabled to participate in AI, rarely are they able to control decisions themselves [123, 487].

In the face of this, collective action presents a way for people to work toward shared goals by banding together. A significant body of existing research examines the power of collective action to bring about positive change and investigates the ways that researchers can support collective action (e.g., [61, 319, 337, 422, 551, 580]). However, there are many unknowns around the best ways to employ collective action within the context of algorithmic systems.

As such, some emerging research has investigated how to empower affected people in their efforts at algorithmic harm remediation. For example, Protective Optimization Technologies (POTs) provide ways for affected parties to “correct, shift, or expose harms” in algorithmic systems without the cooperation of service providers [313]. Data leverage, which presents ways for people to wield their own input data to harm or support technologies and associated companies [552], can occur in the context of algorithmic systems. I orient this work accordingly and consider it in line with the notion of tactics from below in which people “build power over technology when power has not been given” [568].

Part I

Everyday Algorithm Auditing

Chapter 3

Understanding the power of everyday users in surfacing harmful algorithmic behaviors

This chapter is adapted from my paper [474]:

Hong Shen*, Alicia DeVos*, Motahhare Eslami, Kenneth Holstein. “Everyday Algorithm Auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors.” In *Proceedings of the ACM on Human-Computer Interaction* 5. CSCW2 (2021). *indicates equal contribution

A growing body of literature has proposed formal approaches to audit algorithmic systems for biased and harmful behaviors. While formal auditing approaches have been greatly impactful, they often suffer major blindspots, with critical issues surfacing only in the context of everyday use once systems are deployed (Chapter 2). Additionally, as argued in Chapter 6, algorithm audits are currently situated within larger structures of power in limiting ways. Recent years have seen many cases in which *everyday users* of algorithmic systems detect and raise awareness about harmful behaviors that they encounter in the course of their everyday interactions with these systems. However, to date little academic attention has been granted to these bottom-up, user-driven auditing processes. In this chapter, I propose and explore the concept of *everyday algorithm auditing*, a process in which users detect, understand, and interrogate problematic machine behaviors via their day-to-day interactions with algorithmic systems. I argue that everyday users are powerful in surfacing problematic machine behaviors that may elude detection via more centrally-organized forms of auditing, regardless of users’ knowledge about the underlying algorithms. I analyze several real-world cases of everyday algorithm auditing, drawing lessons from these cases for the design of future platforms and tools that facilitate such auditing behaviors. Finally, I discuss work that lies ahead, toward bridging the gaps between formal auditing approaches and the organic auditing behaviors that emerge in everyday use of algorithmic systems.

3.1 Introduction

Algorithmic systems increasingly exercise power over many aspects of our everyday lives, including which advertisements or social media posts we see, the quality of healthcare we receive, the ways we are represented to our peers or potential employers, and which neighborhoods are subjected to increased policing [151, 206, 379,

463, 540]. These systems, however, are not infallible. A growing body of research has drawn attention to the societal impacts of algorithmic systems' behaviors, examining the ways these systems can, whether inadvertently or intentionally, serve to amplify existing biases and social inequities or create new ones (e.g., [91, 379, 512, 597]). In response to these concerns, scholars have proposed a number of approaches to audit algorithmic systems for biased, discriminatory, or otherwise harmful behaviors¹—both internally, within the organizations responsible for developing and maintaining these systems, and externally [417, 447].

Algorithm auditing approaches that have been discussed in the literature often occur outside the context of everyday use of an algorithmic system or platform: auditors, such as researchers, algorithmic experts, and activists, initiate, conduct, or orchestrate the auditing process. While such audits have made significant contributions in detecting biases and harmful algorithmic behaviors, they often fail to surface critical issues. For example, even after devoting months of effort to internal auditing, machine learning development teams often struggle to detect and mitigate harmful biases in their systems due to their own cultural blindspots [256]. Additional reasons why formal auditing approaches can fail to detect serious issues include the presence of unanticipated circumstances or social dynamics in the contexts where a system is used, as well as changing norms and practices around the use of algorithmic systems over time [190]. These factors make detecting certain harmful algorithmic behaviors difficult for auditors unless they are embedded within particular real-world social or cultural contexts [127, 254, 462, 589].

While existing formal audits often struggle to detect critical issues in algorithmic systems, recent years have seen many cases in which *users* of algorithmic systems detect and raise awareness about biased and harmful behaviors that they encounter in the course of their everyday interactions with algorithmic systems. One recent example is the highly publicized case of Twitter's image cropping algorithm exhibiting racial discrimination by focusing on white faces and cropping out Black ones. Twitter users began to spot issues around this algorithm and came together organically to investigate. Through online discussions, they built upon one another's findings to surface similar biases or to present evidence or counter-evidence for a pattern discovered by another person. This occurred even though the company stated that "it had tested the service for bias before it started using it" [249]. Twitter's testing procedures failed to detect this bias because during real world usage, users were interacting with the cropping algorithm in ways the team did not anticipate up front.

Although similar auditing behaviors have been observed around a range of algorithmic systems, including image search engines, machine translation, online rating/review systems, image captioning, and personalized advertising, many open questions remain regarding how users of algorithmic systems come to detect, report, and theorize about harmful algorithmic biases and behaviors in their day-to-day use of a platform. What factors facilitate "successful" user-driven algorithmic audits, and what factors prevent such audits from having an impact? How might we effectively support everyday users' motivation and strengths in surfacing harmful algorithmic behaviors, so as to overcome the limitations of existing auditing approaches?

In this chapter, we take a first step toward answering these questions by proposing and exploring the concept of *everyday algorithm auditing*: a process in which users

¹Throughout this chapter, I distinguish between harmful algorithmic biases and harmful algorithmic behaviors more broadly, as appropriate. Although these categories often overlap, I acknowledge that not all algorithmic harms are best understood as "biases," and not all algorithmic biases are necessarily harmful [74].

detect, interrogate, and understand problematic machine behaviors via their daily interactions with algorithmic systems. In contrast to formal auditing approaches, everyday algorithm auditing occurs in the context of “everyday use” of an algorithmic system or platform. Accordingly, throughout this chapter, we use the term “everyday users” to refer to social actors who take part in everyday use of an algorithmic system. We draw insights from past literature in “everyday resistance” [139, 460] and “algorithmic resistance” [542] to theorize such behaviors. We interpret everyday algorithmic auditing efforts as a form of everyday resistance by which users actively, continuously question and repurpose mass cultural products in their everyday lives to resist the hegemonic culture forms of their times. In resisting algorithmic harms and conducting everyday audits, everyday users usually form counter-publics, “parallel discursive arenas” [186], where communities impacted by harmful algorithmic behaviors come together and participate in their own forms of collective sensemaking, hypothesis development, and algorithm bias detection.

We adopt an exploratory case study approach, examining several recent cases to understand the nature and dynamics of everyday algorithm audits. In particular, we examine two types of algorithmic sociotechnical platforms, the Twitter cropping algorithm and online rating algorithms, that have been the target of everyday algorithm audits in recent years. First, we analyze the Twitter racial bias case mentioned above [249]. Second, we look at another everyday auditing process investigating potential gender bias caused by the same algorithm [411]. Third, we turn our attention to a case surrounding Yelp’s review filtering algorithm, in which a large group of small business owners came together to verify its bias against businesses that do not advertise with Yelp [285]. Finally, we examine a case surrounding Booking.com’s online rating algorithm, wherein users came together to audit the system after noticing that the ratings calculated by the algorithm did not match their expectations [168]. In all of these cases, we compare the different paths everyday users took to detect, examine, discuss, and understand biases in algorithmic systems’ behavior.

Analyzing these cases and comparing their dynamics led us to a process-oriented view of everyday algorithm audits encompassing (1) initiation of an audit to (2) raising awareness of observed issues to (3) hypothesizing about observed behaviors and testing an algorithmic system to ideally (4) some form of remediation. Although some everyday audits may terminate before touching all of these phases, this process-oriented view provides a useful high-level understanding for describing the paths everyday audits can take, comparing different everyday audits, and envisioning new kinds of interventions to facilitate such audits. We argue that day-to-day users are powerful in surfacing problematic machine behaviors that may elude detection via existing, formal auditing approaches, even when users lack technical knowledge of the underlying algorithms. Furthermore, we suggest that everyday algorithm auditing may be most powerful when audits are conducted collectively, through interactive discussions among users with a diverse range of experiences.

Our case analysis and process model of everyday algorithm audits inform a series of design implications for future work toward supporting more effective everyday algorithm auditing. We outline five broad categories of potential design interventions to support everyday audits—a) community guidance, b) expert guidance, c) algorithmic guidance, d) organizational guidance, and e) incentivization—and discuss potential design trade-offs. Finally, we close with a discussion of work that lies ahead, in order to bridge the gap between existing algorithm auditing approaches in academia and industry versus everyday auditing behaviors that emerge in day-to-day use of algorithmic systems.

3.2 Everyday Algorithm Auditing

We develop the concept of *everyday algorithm auditing* as a general framework to capture and theorize a diverse set of emerging user behaviors around algorithmic systems. We define everyday algorithm auditing as the ways everyday users detect, understand, and/or interrogate problematic machine behaviors via their day-to-day interactions with algorithmic systems. Following [447], we use “auditing” to refer to behaviors undertaken, whether formally or informally, to test and investigate whether a given algorithmic system operates in a socially harmful way, such as behaving in ways that produce inequitable outcomes along lines of class, race, or gender. Building on Suchman [504], we argue that everyday algorithm auditing practices are situated actions in which everyday users encounter problematic machine behaviors via their routine interactions with algorithmic systems, then adjust their behaviors to understand and act on what they encountered.

Recent years have seen several cases of everyday algorithm auditing, with some publicized widely. However, apart from some notable exceptions [168, 169], to date little academic attention has been granted to characterizing these bottom-up, user-driven auditing behaviors. In this chapter, we aim to bridge the gap between algorithmic auditing approaches proposed in the academic research literature and auditing behaviors that users exhibit day-to-day. In doing so, we explore what lessons can be drawn from the study of everyday algorithm auditing to overcome limitations of existing auditing approaches in the literature.

In this section, we first describe the scope and boundaries of everyday algorithm auditing. We then elaborate on how we develop the concept based on two streams of past literature, namely, everyday algorithmic resistance and counterpublics. As discussed below, we argue that everyday algorithm auditing can be viewed as a form of *everyday resistance*. When performed by a group of users, everyday algorithm auditing can additionally be understood through the lens of *counterpublics*, where members of often disadvantaged and marginalized social groups participate in their own form of collective sensemaking, opinion formation, and consensus building.

3.2.1 Scope and Boundaries

As an emerging phenomenon, the conceptual boundaries of everyday algorithm auditing remain somewhat fluid. In this chapter, we scope this concept broadly, to include any actions taken by everyday users to notice, interpret, question, or bring attention to problematic algorithmic behaviors. Below we highlight some key aspects of the concept to help elucidate its scope: (1) what algorithmic expertise the everyday users have, (2) the extent to which an everyday audit represents a collective effort, and (3) how organic the entire process is. These three dimensions, reviewed below, are not intended to be exhaustive; rather, they serve as a starting point to better illuminate the definition.

Algorithmic Expertise: While everyday algorithm auditing is defined by everyday use, this does not imply that the users engaging in these behaviors necessarily lack technological skills and algorithmic expertise. Sometimes, everyday audits are conducted by users who have a high degree of relevant technical knowledge. For example, Sweeney, who is a professor of the practice of government and technology, performed her classic audit of online ad delivery only after she found that her own name was linked with an arrest record during a casual search on Google [512]. In this case, Sweeney acted as a situated user while also having relevant technical expertise in the area. In another example, Noble, a professor of information studies

and expert in algorithmic bias and discrimination, began her examination of racist bias in search engines after trying to find an engaging activity for her stepdaughter and nieces by googling “black girls.” Instead, her web search yielded pornography [379]. In a third example, Buolamwini, a researcher with expertise in algorithmic bias and accountability, embarked on the Gender Shades project to formally audit facial recognition systems after facial analysis software that she needed to use for an engineering project would not recognize her face unless she wore a white mask [321].

In other cases, everyday users may have little algorithmic expertise. For example, many small business owners collectively initiated an everyday audit of Yelp’s algorithm after noticing that positive reviews were not showing prominently on their business pages [169]. As another example, a group of Google Maps users uncovered that searching the N-word directed them to the White House, which Obama then occupied [191]. Since “expertise” is multi-faceted, everyday users may have some relevant technical expertise, while lacking specific knowledge of machine learning.

Collectiveness: Everyday algorithm audits can also vary in the extent to which users work together collectively. Some cases are more individually led, such as the case when Alciné discovered that Google’s image recognition algorithm labelled some Black people as “gorillas” [231]. Similarly, Sweeney conducted her audit alone, a highly individual case [512]. Other cases, such as the time when LGBTQ+ YouTubers came together to investigate the demonetization of their videos, are more collective [433]. It is worth noting that having a large number of users involved does not always mean an everyday audit is collective. For example, in the case where Booking.com users instigated an audit [168], they did so with little ability to communicate amongst themselves, resulting in a less collective audit despite its large scale.

Organicness: Though everyday algorithm auditing is based on everyday use, in practice the process can be more or less organic. Some everyday audits begin with an outside intervention but become more organic and user-led over time. For example, ImageNet Roulette was a simple online tool built by artists and researchers to support users in exploring and interrogating the input/output space of an image captioning model trained on the ImageNet dataset. However, after the tool was released to the public, users had autonomy in deciding how to use it. Collective auditing behaviors emerged organically through discussions on social media, as users shared findings and hypotheses and sometimes built upon each other’s explorations [130, 363]. By contrast, other everyday audits start organically but later turn highly organized. For example, Sweeney, Noble, and Buolamwini all commenced their everyday audits inadvertently through everyday use, then turned to more formal auditing methods for testing and analysis [90, 379, 512]. Still other everyday audits are organic throughout the entire process. In the case of racial bias in Twitter’s cropping algorithms, Twitter users started and self-directed the whole audit with minimal intervention (in the form of a few tweets from Twitter employees whose input was invalidated by the users’ auditing process [116, 359, 559]).

3.2.2 Everyday Algorithmic Resistance

We argue that everyday algorithm auditing can be thought of as a form of *everyday algorithmic resistance* [542]. In their seminal work, Certeau and Scott [139, 460] developed the idea of “everyday resistance” to examine the ways in which people exercise their agency in front of dominant power structures and hegemonic culture formats in their everyday lives. Instead of thinking of resistance as organized actions that pose a revolutionary challenge, they foreground a different type of resistance that is

less visible and more incidental but nevertheless constantly contesting the existing power relations in everyday forms.

In the algorithmic era, we see similar resistance by everyday users around algorithmic systems. For example, in his analysis of gay men's resistance of Blued, China's largest gay dating app, Wang looked how users act on the data they provide to dating apps to shape algorithmic dating outcomes in a different direction than what the system was originally designed for [557]. Similarly, Devito et al. examined how Twitter users mobilized via hashtag to resist the introduction of algorithmic curation of Twitter's timeline [146]. Instead of characterizing the relationship between users and algorithms as passive consumers versus omnipotent machines, this line of research helps us foreground users' agency, capacity, and tactics in their regular interactions and contestations with algorithms.

As a form of algorithmic resistance, in everyday algorithm auditing we see how users constantly test the limits of algorithms in their everyday use of the system. Sometimes, these auditing behaviors might be incidental and incremental: for example, a user might encounter harmfully biased algorithmic outputs and merely report their observations online. In other cases, these incidental and incremental behaviors might spur a group of users to work together, build on each other's findings, and collectively interrogate the underlying system.

3.2.3 Counterpublics

We argue that everyday algorithm auditing, when performed by a group of users collectively, can also be viewed as a form of *counterpublics*. Nancy Fraser developed this influential idea as a critique to the concept of a universal public sphere developed by Habermas [232]. She proposes that counterpublics can be understood as "parallel discursive arenas" [186], where members of often disadvantaged and marginalized social groups come together and participate in their own form of collective sensemaking, opinion formation, and consensus building.

In the past a few years, we have seen growing efforts in building counterpublics around algorithmic systems. For example, Geiger examined bot-based collective blocklists in Twitter, which have been developed by volunteers to combat harassment in the social networking site [199]. Xiao et al. [583] discussed how finstas—secondary Instagram accounts that feature content often unacceptable to users' primary accounts—have developed forms of counterpublics for young people to share emotional or vulnerable content with their close friends.

In everyday algorithm auditing, we saw similar counterpublics emerging. Indeed, many influential cases started with individual users complaining about certain socially harmful machine results in their everyday lives, but ended up with a large group of users working together for a collective and collaborative action. For example, in 2019 a group of YouTubers came together to show that the YouTube recommendation algorithm demonetizes LGBTQ+ content, resulting in a huge loss of advertising revenue for LGBTQ+ content creators [433]. They tested the system with new examples, validated or invalidated each other's findings, and interdependently probed the possible working mechanisms of the underlying system. During the process, they formed temporary counterpublics against the dominant algorithm, collectively making sense of the system, forming their own understanding and opinion, and building consensus toward possible remediations. In contrast to crowdworkers independently working on decomposed subtasks determined entirely by outside parties, in everyday algorithm auditing, the crowd collectively decides the course an audit takes, often through discussion threads or other social channels. For example,

in the YouTube LGBTQ+ demonetization case, users started forming counterpublics on different platforms such as content creation websites (e.g., Patreon) and social media (e.g., Twitter) to work together detecting algorithmic biases as well as supporting affected communities.

When forming counterpublics, users are involved in a process of *collective sense-making* [564] of observed algorithmic behaviors. A rich tradition in HCI and CSCW has contributed to our understanding of how groups of people can work together, build upon each other's explorations and insights, and work to achieve common goals. Past work has examined collective sensemaking across a range of domains, including citizen science [128], knowledge mapping and curation [179, 200, 234], and social commerce [111], to name just a few. Similarly, in everyday algorithm audits, users come together to collectively question, detect, hypothesize, and theorize about problematic machine behaviors during their interactions with algorithmic systems.

3.3 Methods

We set out to understand what we can learn from existing everyday algorithm audits in order to inform the design of platforms and tools that can better support these practices. Thus, we asked: How can we better understand the characteristics, dynamics, and progression of everyday auditing practices? How can we support everyday users in detecting, reporting, and theorizing about problematic machine behaviors during the course of their interactions with algorithmic systems?

In order to understand the nascent phenomenon of everyday algorithm auditing, we adopted an exploratory case study approach [410], examining several recent, prominent cases. We began with a small set of high-profile cases that at least one of the co-authors was already familiar with. We iteratively reviewed and discussed these known cases to form the initial idea of "everyday algorithm auditing". Through this process, we generated related keywords (e.g., "auditing," "testing," "algorithm," "algorithmic," "bias," "harm," "users," "collective," and "community"). Next, we searched news media (e.g., Google News, Google Search) and social media (e.g., Twitter, Facebook) using combinations of these keywords to get a rough sense for the scope of this phenomenon. The search yielded 15 cases (see Table 1) that met our definition for everyday algorithm auditing: one or more everyday users act to detect, understand, and/or interrogate biased and harmful algorithmic behaviors through their everyday use of a platform. These cases span multiple algorithmic domains, including image captioning, image search, machine translation, online rating/review, image cropping, credit scoring, and advertising and recommendation systems. We reference these cases throughout the chapter to ground our discussion in concrete examples. However, this set of cases is by no means comprehensive. Furthermore, as discussed in our *Design Implications* section, we expect that these existing cases represent only a thin slice of what is possible for everyday algorithm audits in the future. At this exploratory stage, our aim is to begin formalizing this concept, to help guide future empirical and design research in this space.

Case Selection: To support depth as well as breadth in our analysis of everyday algorithm audits, we chose a small subset of four cases to examine in greater detail. To select these cases, we iteratively extracted patterns from our initial dataset through a series of discussions among our research team. We set out to choose a set of cases that a) span multiple domains and vary along the three dimensions outlined in the *Scope and Boundaries* section (algorithmic expertise, collectiveness, organ-icness) and b) were accessible to us via multiple data sources (e.g., user discussion

Domains	Cases	Descriptions
Search	Google Image Search [379]	Researcher Noble searched “black girls” on Google and found out the results were primarily associated with pornography.
Rating/review	Yelp advertising bias [169]	Many small business owners on Yelp came together to investigate Yelp’s potential bias against businesses that do not advertise with Yelp.
	Booking.com quality bias [168]	A group of users on Booking.com scrutinized its rating algorithm after realizing the ratings appeared mismatched with their expectations.
Image cropping	Twitter racial cropping [528]	Researcher Madland posted an image on Twitter of himself and a Black colleague who had been erased from a Zoom call after Zoom’s algorithm failed to recognise his face. Twitter automatically cropped the image to only show Madland. After Madland posted this too on Twitter, users joined him in testing the Twitter cropping algorithm.
	Twitter gender cropping [411]	A researcher noticed that when VentureBeat shared an article with an image of two women and two men on Twitter, the algorithm cropped out both women’s heads. This was followed by other reports and testing by Twitter users.
Image captioning	ImageNet Roulette [130]	Artists/researchers Crawford and Paglen built an interface to help users test a model trained on the ImageNet dataset. Although technical experts provided a tool that supported everyday auditing behaviors, users had autonomy in directing their efforts and using images of their choosing. Collective auditing behaviors emerged through discussions on social media.
Image recognition	Google Photos [231]	Programmer Alciné uploaded photos to Google Photos and noticed the app labelled his Black friends as “gorillas.”
Advertising	Google’s online ad delivery [512]	Researcher Sweeney performed her classic auditing of online ad delivery after she found her own name presented as linked with an arrest record in one of her many typical searches on Google.
Recommendation systems	YouTube LGBTQ+ demonetization [433]	A group of YouTubers found that the YouTube recommendation algorithm demonetizes LGBTQ+ content, resulting in a huge loss of advertising revenue for LGBTQ+ content creators.
	Google Maps [191]	A group of users reported that when they searched for the N-word on Google Maps, it directed them to the Capitol building, the White House, and Howard University, a historically Black institution. Other users joined the effort and uncovered other errors.
	TikTok recommendation algorithm [288, 485]	A group of users found that TikTok’s “For You Page” algorithm suppresses content created by people of certain social identities, including LGBTQ+ users and people of color. As a result, they worked together to amplify the suppressed content.
Translation	Google Translate Quality-of-service harm [427]	Computer engineer Rezasab noticed that Google Translate mistranslated “condolences” in Persian to “congratulations” in English when the phrase was directed toward people of certain countries, like Lebanon. In response, other users collectively tested the system using their own examples, showing this bias for some other nations, and discussed their observations.
	Google Translate gender bias [387]	Social media users tested Google Translate for gender bias after noticing it associates certain genders with professions and activities when translating from gender-neutral languages such as Turkish, Hungarian, and Farsi.
Credit Card	Apple Card [546]	Tech entrepreneur Hansson noticed the credit limit of his Apple Card was 20 times higher than his wife’s, even though she has a higher credit score and they file joint tax returns.
Facial recognition	Gender Shades [90, 91]	Researcher Buolamwini noticed a problem when she was working with a facial analysis software: the software didn’t detect her own face. She went on to evaluate the accuracy of a number of AI-powered gender classification products on people of different genders and skin tones.

Table 3.1: 15 cases of everyday algorithm auditing across various domains, which we use to ground our discussion throughout this paper. Bold cases are analyzed in depth.

posts, news articles, research studies), enabling a rich analysis.

In particular, we chose two different domains that each have each been the target of everyday algorithm audits in recent years: image cropping algorithms and rating platform algorithms. Within each domain, we examined two cases that vary across

the three dimensions discussed above to support comparison. Figure 3.1 summarizes the four cases, visualizing varying degrees along each dimension. For example, there are highly collaborative, less collaborative, and individual levels of collectiveness among the four cases. In addition, these cases span multiple levels of algorithmic expertise and organicness. The cases serve as illustrative examples to better understand how everyday audits can work, what paths they can take, and what impacts they can have.

For each of our four primary cases, we gathered and reviewed discussion threads from the platforms where the everyday audits took place, then supplemented this information by drawing upon relevant media (e.g., news articles, academic publications) collected in our earlier search. We describe these four cases, their characteristics, and their dynamics in the next section. From this process of examining the progression of each case, a set of broader lifetime dynamics emerged, which we describe later in the chapter.

3.4 Case Studies

Now, we focus on four cases of everyday algorithm auditing to illustrate and investigate the characteristics and dynamics of these audits more broadly. We look at two algorithmic sociotechnical platforms that each have been the target of everyday algorithm auditing in recent years. For each platform, we examine two different categories of biases that were detected via everyday algorithm auditing, as well as the different paths users took to conduct these audits. We compare these cases with each other to understand similarities and differences between everyday audits, and we consider how the paths taken contribute to their impacts.

3.4.1 Twitter Image Cropping Algorithm

Twitter employs algorithms throughout its infrastructure to make decisions about how content is presented to users. One such algorithm on Twitter uses neural networks to automatically crop the images users tweet, with the aim of improving users' experiences [521]. Twitter crops images included in tweets based on how many images appear in a tweet, where more images contained in a tweet mean that each image appears as a smaller thumbnail; the dimensions for this cropping have been shared widely among Twitter users trying to understand why images in their tweets appear the way they do [144, 367]. Twitter's cropping algorithm attempts to find the parts of images most interesting to people—as judged by how likely people are to look at a certain region and termed “saliency”—and center that part of the image within the frame, theoretically cropping the least interesting parts of the image out of the thumbnail [521]. This appears to be a good solution for users and Twitter alike: maintain consistent design of tweets while highlighting the part of an image users most want to see.

But Twitter users noticed that the image cropping did not always match what they would have guessed to be the most “salient” parts of images: specifically, they noticed problematic trends in the behavior of the automated cropping, leading them to suspect bias within the cropping algorithm. In one high-profile case in September 2020, Twitter users investigated potential racial bias in the way the cropping algorithm decided whom to focus on and whom to crop out. In another case that gained less traction in January 2019, Twitter users considered possible gender bias in the

Cases	Algorithmic Expertise	Collectiveness	Organicness
Case 1: Twitter Racial Bias	Many users had some technical expertise, but few had relevant algorithmic expertise.	Users collaborated to discuss, theorize, strategize, and test on the platform.	The audit was almost entirely organic, though a few of the platform's employees stepped in once the audit gathered attention, and attempted to help.
Case 2: Twitter Gender Bias	Some users were computer science researchers with relevant algorithmic expertise.	Users acted mostly individually; participation rarely involved discussion.	This audit was entirely organic.
Case 3: Yelp Advertising Bias	Users did not necessarily have algorithmic expertise beyond using the platform to advertise their businesses.	Users came together to discuss, theorize, and strategize on the platform's forums.	This audit was almost entirely organic, though the resulting lawsuits were certainly organized.
Case 4: Booking.com Quality Bias	Users did not necessarily have algorithmic expertise beyond using the platform to review hotels. Later, a researcher, who also stumbled upon this bias via everyday use, conducted a formal audit to establish firmer evidence for the bias.	Users acted mostly individually, as there were few mechanisms for communication on the platform.	The audit was entirely organic.

FIGURE 3.1: High-level descriptions of our four primary case studies across the three dimensions outlined in the *Scope and Boundaries* section. These cases serve as illustrative examples to better understand how everyday audits can work, what paths they can take, and what impacts they can have. Darker shades represent higher levels along a given dimension; lighter shades represent lower. Note that less variation is presented along the “Organicness” dimension given that this chapter focuses on everyday audits toward the more organic end of the spectrum as less organic everyday audits have been well described by past literature.

cropping algorithm’s decisions. Below, we describe these two cases and the paths they took along with their dynamics, results, and impacts.

Case 1. Racial Bias:

Our first case began when a user of the video conferencing platform Zoom realized that a colleague’s trouble using virtual backgrounds might have to do with the software not recognizing darker skin, then tweeted about it [344]. The user tweeted images with the explanation, single shots that captured the Zoom screens of both the colleague, who has darker skin, and the tweeting user, who has lighter skin, side-by-side. The final tweet cropped this wide image to show a thumbnail of just the lighter-skinned person, prompting the tweeter to notice unexpected behavior and theorize that Twitter’s cropping algorithm might be racially biased, then tweet about this new subject [346]. The individual tweeting—that is, the initiator of the everyday audit—stumbled upon Twitter’s photo cropping issue in the midst of regular use of Twitter: sharing information with others on the platform. Incidentally, the information the initiator set out to share on Twitter was related to the issue noticed on Twitter itself, which may have helped the user enter the correct headspace for more awareness of these types of strange and potentially biased behaviors.

Having developed a hypothesis that the cropping issues were related to skin color, the initiator kicked off testing by tweeting the same image as before but flipped so that the person initially on the right was now on the left and vice versa [345]. This test also showcased the lighter-skinned person. Others on Twitter began to join in, tweeting images of their own creation to see what the algorithm would prioritize and what the algorithm would crop out. At the same time as users tested whether the algorithm was indeed biased by skin tone, other users developed new hypotheses that would lead to more tests in attempts to confirm or invalidate.

For example, one highly shared thread put Mitch McConnell at one end of a long image and Barack Obama at the other, finding that McConnell appeared in the thumbnail no matter which end he appeared on [34]. When a new hypothesis surfaced suggesting that the preference in this case might be based on the clothing, specifically the ties, that the two politicians wore in the photos used—perhaps the algorithm prioritizes the color red, which is worn by McConnell while Obama wore a blue tie—a new test was run: exactly the same as before, but switching the ties worn by the two men. McConnell again appeared in both thumbnails.

Other theories arose, leading many users to conduct a wide range of tests in tweets involving elements such as background shade [115, 371], background color [30], cartoon characters [348, 482], lighting [390], women [391], different numbers of people [1], a variety of skin tones [31, 78], image contrast [372, 373], and even dogs [26, 350]. Throughout, users built on and countered each other's suppositions, wielding the tests they had run as evidence supporting or opposing the claims of various hypotheses.

The audit by everyday Twitter users of Twitter's cropping algorithm for racial bias had both specific and broader impacts. Broadly, the highly collective nature of this audit where users worked together to determine a direction for the audit, constantly interacting with each other along the way, led to the creation of a large counterpublic space in which thousands of Twitter users interacted [34]. This space allowed users to question the algorithmic authority of the very platform they conversed on, develop concrete concerns, and push for change. As might be expected for a group this large, the media picked up on what was happening and many articles were published about the issues highlighted by the audit (e.g., [86, 279, 448, 528]). In addition to—or perhaps because of—these broad reverberations, Twitter provided an official response addressing how they would attempt to fix the issues highlighted by the everyday audit: ultimately, they aimed to reduce cropping's reliance on an ML model and allow for more user controls as well as ensure that image thumbnails in the preview of a tweet accurately represented what the image thumbnails would look like once the tweet was live [8].

Case 2. Gender Bias:

The second case began when VentureBeat shared an article with an image of two women and two men on Twitter [54]. A Twitter user pointed out that both women's heads had been cropped out while neither of the men's had. As a result, while the cropping algorithm highlighted the men's heads, it focused instead on women's chests [411]. The individual who shared this issue to start the everyday audit had noticed it while looking at the VentureBeat tweet, a normal usage of the Twitter platform. Coincidentally—ironically?—the VentureBeat article featured four prominent AI researchers, with one focusing on responsible AI [278], so this may have nudged the user into a mindset that took notice of the cropping issue.

Unlike the previous, racial cropping issue, very few other Twitter users joined this everyday audit. Because of this, there was little additional hypothesis forming or testing, though some conversation did happen in replies to the initiating tweet [412] as well as in a thread begun by another user that mostly focused on whether Twitter or VentureBeat was to blame [196]. The gendered cropping issue was revisited in September 2020 and received more (but still relatively little) engagement, including minimal testing of how Twitter crops other images, on the heels of the prominent racial cropping case discussed above.

Overall, the gendered cropping issue gained little traction. It stayed more individualized: only a handful of people participated when it was first brought up in 2019, and though over a thousand liked the 2020 tweet resurfacing the issue, still very few engaged in any sort of conversation about the issue. One might suspect that the initiators of the gender bias case were not as influential or active as the initiators of the racial bias case; but the initiators of both the racial cropping audit and the gender cropping audit have comparably large Twitter followings. Examining this under the lens of meme virality produces some potential explanations for why this might have occurred. While both Twitter cases can be understood as a meme in the sense that it is a “unit of cultural transmission” [138], this one lacks the elements of mimicry and remix some view as vital to meme-hood and virality [476]. In the previous case of racially biased cropping, mimicry and remix can be seen as users build upon each other’s observations, hypotheses, and testing strategies in the form of tweets and retweets. Another possible explanation for the disparity in engagement is that current events in 2020, such as the ubiquity of the Black Lives Matter movement, might have led people to be more aware of and more interested in issues of racial justice. This is a particularly appealing possibility given that information in memes tends to propagate from mainstream news to social media [324]. Since content with greater emotive strength is more likely to spread [226], another possibility is that Twitter users are less interested in gender-related bias than race-related bias. However, in November 2020, a Twitter user apparently unconnected to either of the past conversations around gender cropping tested gender bias in a thread that received almost four times the amount of engagement as gender cropping before, suggesting that people are interested in the gender cropping issue as well [156]. So, why did these two types of everyday audits, both occurring on the same platform due to biases introduced by the same algorithm, receive different levels of attention and take diverging paths? This remains an open question that needs further investigation.

Likely because of the small number of people involved in the gender cropping audit, it had little impact. A small counterpublic space did arise, briefly, but not much else occurred. The small number of people kept the audit mostly individualized in nature, with people who engaged mostly supporting the audit’s existence rather than participating or helping to direct it. Not many people meant that the audit had much more difficulty gaining any sort of traction beyond itself and was ultimately unable to garner publicity or platform change like the racial cropping audit.

3.4.2 Rating Platform Algorithms

Rating platforms are another type of sociotechnical system that use algorithms to determine how to present information to users. Final ratings can have huge impacts on a business’s livelihood as consumers look to the ratings to direct their behaviors: for instance, restaurants were 30–49% more likely to fill their seats following just a half-star rating increase on Yelp [29]. Though some users might assume that a final rating is determined as a simple, raw average of all the ratings given by individual users, many rating platforms do not follow this method. For example, in calculating a product’s rating, Amazon considers factors including how old the associated review is, whether the reviews have verified purchases, and the number of votes the review received for being helpful [73]. Other platforms like Yelp [586] do follow this method of straightforward averaging but only include authentic reviews, as classified by their algorithms.

Additionally, the mechanisms behind these rating systems often lack transparency [169], which has caused controversy about algorithmic bias in recent years. In May 2016, for example, Uber was accused of exploiting its rating system's opacity to slowly decrease the ratings of its Australian drivers so that it could suspend them and charge high reinstatement fees [525]. Similarly, Yelp has been criticized for lack of transparency around the inner workings of its rating system [184]. After suspecting that the platform manipulated its rating system in order to force businesses to pay for advertising, Yelp users conducted an investigation. Below, we explore the Yelp everyday algorithm auditing case as well as the Booking.com case in which users scrutinized the rating algorithm after realizing the ratings appeared mismatched with their expectations.

Case 3. Yelp Advertising Bias:

The Yelp case has a long history, beginning as early as 2011 when small business owners on Yelp began to suspect that Yelp manipulated its review filtering algorithm to hide positive reviews [44], which would then compel business owners to pay for advertising to boost their rating positions to make up for having fewer positive reviews. Many small business owners on Yelp claimed to have noticed this apparent algorithmic aberration after they had received phone calls from the Yelp sales team about advertising with Yelp; when they rejected the offer, opting not to pay for Yelp advertising, the positive reviews for their businesses started to get filtered out. Even worse, the negative reviews for their businesses inexplicably moved to the recommended page. The business owners took, naturally, to Yelp, where they began posting on the Yelp Talk discussion forums about what they had noticed (e.g., [44]). In this way, the business owners together initiated the everyday algorithm audit by sharing the perceived issue with others after encountering it through typical usage of the platform.

Having developed a hypothesis that the filtering issues were related to the advertising offer, business owners began to test by seeing what happened to reviews for businesses following calls from Yelp Sales agents (e.g., [134]). Results of these tests were circulated to others via the discussion forums, where business reviewers joined business owners in conversation about the filtering algorithm. Alternate hypotheses arose, such as the theory that reviews were being filtered out because writers appeared illegitimate based on profile appearance or behavior—e.g., no profile picture and few friends; writing too few, too pithy, too extreme, or too infrequent reviews [175, 440]. These conjectures were often based on users' folk theories about how the filtering algorithm operated and not on concrete data [169]. Yelp users probed these hypotheses through debate in the forum and through further testing that often involved business owners analyzing the elements of filtered and recommended reviews and business reviewers visiting establishments and leaving reviews to test whether they would be visible [393]. Business owners participated especially intentionally, as they often had personal motivation if the algorithm had started filtering their business's positive reviews or recommending their business's negative reviews. Throughout, the users referred to the tests they had run and witnessed as evidence in support of or in opposition to various proposed hypotheses.

The audit of Yelp's review filtering algorithm by everyday Yelp users has a number of ramifications. In perhaps the largest and most visible result, the claims of the audit led to many lawsuits filed over the years with almost 700 accumulated over time [286, 395]. Though these lawsuits were ultimately unsuccessful [209], they

created a large amount of public awareness and forced Yelp to engage with the complaints, at least in the legal arena. Perhaps the widespread initiation of this everyday audit, with different business owners in widely separated locations noticing the inconsistency in the filtering algorithm's behavior at the same time in a sort of collective initiation, forged the way for such large impact. In a smaller but still important way, the everyday audit created awareness and support in a counterpublic space within the Yelp platform on Yelp Talk discussion forums as users collectively considered the filtering algorithm. Yelp unintentionally allowed this, but it formed an imperative aspect of this everyday audit and highlights the value of providing an ecosystem for communication and discussion between audit participants.

Case 4. Booking.com Quality Bias:

The other case of rating platform bias began when Booking.com users rated hotels where they had stayed and noticed a discrepancy between their desired total rating and what the algorithm calculated. Booking.com asks users to assess a number of different elements of their lodging experience, then uses an algorithm to calculate an overall rating based on this assessment. The discrepancy appeared when disgruntled hotel guests selected the lowest option for every element, aiming for a score of 1 or 0 out of 10, only to discover that the algorithm would produce a score of 2.5 [168]. Similar to the previous, Yelp case, many users realized this strange Booking.com behavior around the same time, initiating the audit by pointing out the issue in the body of their hotel reviews, and they did so during typical platform usage: trying to leave poor reviews.

Unlike the Yelp case, this everyday algorithm audit stayed highly individualized. After Booking.com users reported the issue they observed through reviews, there was no on-platform way for the users to communicate with each other, in contrast to the Yelp case where users could talk on the Yelp platform. Booking.com users thus individually set out to test their hypothesis that the rating algorithm would not dip below 2.5 by trying different combinations of rating inputs. Multiple hypotheses for the Booking.com behavior eventually did emerge, such as that the algorithm was purposefully skewing the ratings to present a better image for low-to-medium quality hotels, that the algorithm rated overly high in general, and that the algorithm was just incorrect [168]. Users attempted to warn others of both the bad hotel experience and the unrepresentative algorithm behaviors, using review writing as a method for raising awareness since this was the only way to communicate on the platform.

In terms of impact, this everyday audit had few. The Booking.com case generated little publicity, despite having proven existing bias. This is especially interesting when taken in conjunction with the Yelp case, which generated an abundance of publicity despite never having proven bias—in fact, the Yelp case was ultimately dismissed due to the lack of scientific evidence [209]. Perhaps this difference arises from a difference in personal injury: in the Yelp case, the livelihoods of many business-owning everyday auditors were at stake, whereas in the Booking.com case, the most at stake for the everyday auditors was the ability to fully trust the ratings upon which they based hotel reservation decisions. Or perhaps more collective everyday audits invite publicity, as the collective nature means that not only are more people aware of the audit occurring, but also they are more aware of it as they actively participate in sensemaking discussions and actions. While Booking.com's audit participants lacked collectiveness, they still changed their behaviors to try to make changes to the system; that is, they signaled to other users a more accurate presentation of their hotel stays, and they trusted the overall platform ratings less

[168]. Booking.com did eventually make changes to its rating system to make the lowest possible score to 1 (instead of 2.5) and allow users to select an overall score that will appear unchanged by an algorithm [81], which are likely indirect results of this everyday audit.

3.5 The Lifetime and Dynamics of an Everyday Audit

Below we propose a process-oriented view of everyday algorithm audits, with phases ranging from (1) initiation of an audit to (2) raising awareness of observed issues to (3) hypothesizing about observed behaviors and testing an algorithmic system to (4) some form of remediation. In practice audits may follow a non-linear path through these phases, and may terminate before reaching all phases, for various reasons discussed below. Regardless, these provide a useful high-level understanding for describing the paths everyday audits can take, comparing different everyday audits, and envisioning interventions to facilitate such audits.

3.5.1 Initiation

Everyday algorithm auditing usually starts in the moment that an individual or group of individuals finds an instance of harmfully biased behaviors through normal usage of an algorithm. Some of the cases such as the Twitter cropping algorithm's racial and gender bias start with one individual user, the *initiator*, and then continue via other users. On the other hand, in some cases such as Yelp's filtering algorithm or Booking.com's rating algorithm, many users notice a potential bias around the same time in different locations with or without previous connections with each other. These behaviors, as we call individual and collective initiation respectively, could play a role in the spread and impacts of an everyday audit.

Everyday audits are often initiated by *incidental* exposure to problematic machine behaviors: in other words, users may not necessarily intend to conduct an audit from the start. Rather, users often stumble upon concerning algorithmic behaviors in the midst of their day-to-day use of a system. For example, in the Twitter racial cropping case began when, in the course of tweeting an image, noticed that out of two people shown in the image, the person with darker skin was excluded from preview thumbnails. While everyday audits frequently begin through incidental exposure, users may be more likely to detect and spread awareness of problematic machine behaviors when they are already "primed" to look for such behaviors. For instance, in the Twitter racial cropping case, the audit initiator noticed a concerning cropping behavior in the context of tweeting about a different racial bias observed on another platform (Zoom). As might be expected, unintentional initiation is more likely to arise from the users' situated actions whereas more intentional initiation is more likely to arise from the actions of external auditors and not in the context of everyday use.

3.5.2 Awareness Raising

After detecting a problematic algorithmic behavior, the initiator or initiators broadcast what has been discovered to others. For example, in both Twitter cropping cases, this took the form of a tweet to other users. Others can then spread the word to even more people. Promoting the audit is a valuable part of the process, as it can bring more people into the discussion and increases visibility of the issue at hand. Sometimes, like in the racial cropping Twitter case, this promotion by other users

creates the opportunity for further sharing and for participating in other parts of the audit like hypothesis forming and testing, which we explain in the next section. Other times, when discussion between users is more limited by a platform like in the Booking.com case where users had to communicate through reviews, this promotion by other users solely creates more awareness among those who see it, helping them to adjust their own behavior in using the system.

Raising awareness has ramifications both internal and external to the everyday algorithm audit. Internally, sharing information about the audit, for example in the racial cropping case on Twitter when some people created larger threads combining information from multiple places, serves to draw more participants into the audit in a cascading, self-referential way that potentially increases its collective nature. That is, the more users promote the audit, the more new people notice its happening and are then more likely to promote the audit themselves. Externally, sharing information about the audit makes it more likely that the awareness will eventually extend beyond the audit, as will be discussed more in Remediation below.

Everyday audit participants also can generate awareness through supporting the mechanisms of the audit. The LGBTQ+ YouTube demonetization is an example of this, as some YouTubers worked together to collect keywords, and a large number of people did a lot of YouTube tagging as well as shaming the company on Twitter for its actions. While this can promote the audit itself, these actions are more focused on assisting—but not guiding—other parts of the audit. Often a core group of people collaborate in the auditing process and are surrounded by a larger ring of people supporting their actions. The lack of many people in this support ring might help explain the lower traction for the Twitter gender cropping case as compared to the racial cropping case. Both the core group and the surrounding ring form a counterpublic space that supports the everyday audit as it progresses.

3.5.3 Hypothesizing & Testing

Following initiation and/or awareness raising, an everyday algorithm audit may progress toward users forming hypotheses for the observed behavior and intentionally testing an algorithmic system to collect further evidence. This process may involve just the individual who initiated the audit or may extend to many users if the initiator raised awareness and inspired others to get involved. Individual users who proceed alone continue with little support either from other situated users or from external experts. In more collective everyday algorithm audits, a group of everyday users organizes and builds on each others' efforts in the process. The users collectively decide on and steer the process, often through online discussion.

When users start hypothesizing about the potential biases of a system and testing their hypotheses, they usually develop “folk theories”: non-authoritative theories a user forms to explain and understand how a technological system works [145, 165]. These theories, regardless of their validity, shape user behaviors and inform how they act around a system. In an everyday algorithm audit, these theories help users to form hypotheses about the sources of a bias, then test them. Sometimes these tests fail, such as switching the tie colors of McConnell and Obama in the Twitter racial cropping algorithm case to test the hypothesis that the algorithm might prioritize red over blue [34]; other times these tests succeed, such as manipulating the algorithm inputs in the Booking.com case to test the hypothesis that the algorithm skewed the minimum score up [168]. These examples show the importance of folk theory development in the process of an everyday audit: the more the platform allows users to develop and test theories, the more users have the opportunity to discover biases.

One way to empower users in this process is *seamful* design; as Eslami et. al. described, adding visible hints disclosing aspects of an automation process, or seams, into the design of an algorithmic system can help users construct better conceptual understandings and theories about the system [165]. This increased transparency can provide users with more visibility, which may aid in detecting potential biases. The increasing number of everyday algorithm audits, along with new regulations around algorithmic bias [212, 261], can push algorithmic sociotechnical systems for a more transparent and seamful design.

Hypothesizing and testing often involves different degrees and forms of collaboration. In cases where users perform the auditing in an individual endeavor, they might spot an issue and quickly report it via social media channels or the platforms' internal discussion forum or, if coming from a background of high technical expertise, they might conduct the auditing in a more systematic way alone. In other more collective cases, users come together to make sense of how the system actually produces such biased results. During this process, they might raise new evidence, test each other's evidence, and collectively search for patterns that emerged from their discussion. For example, in Yelp's filtering algorithm case, users collectively worked together on the Yelp platform itself to form different hypotheses about how the algorithm works, test those hypotheses, and report the results back to the community.

Similar to the motivations behind initiation of the audit, the hypothesizing and testing process can range from organized to organic. More organized testing often involves external stakeholders (e.g., researchers, developers, journalists) who intervene in the process, while more organic testing often arises naturally without outside intervention as everyday users detect and assess algorithmic behaviors and play an active role in determining the course the audit takes. We argue that everyday algorithm auditing can receive—and often benefit from—different degrees and forms of external intervention. For instance, in some cases, developers with insight into the inner workings of a given system may be able to provide useful feedback to guide hypothesis formation, such as in the Twitter cropping case when an engineering team lead chimed in to let audit participants know that facial recognition is not part of the cropping algorithm [559]. To the extent that external interventions represent *guidance* for an otherwise collectively-led audit, rather than directing or orchestrating the collective's activities as in a crowdsourced/collective audit, we do not view such interventions as invalidating the *everyday* nature of the audit.

3.5.4 Remediation

At its core, an everyday algorithm audit has a singular objective: instigate change based on the issues identified and investigated. Change can occur at multiple levels. One type of change creates or increases external awareness through publicity. As the audit gains traction, news and other media sources are likely to become aware of the audit as well and may publish articles on the audit and the issues it highlights. This leads to broad social impact via greater awareness of the issues in the public sphere by all types of people, who may then choose to share this information with still others, alter their own patterns of behavior, or even take part in the audit if it is still occurring.

Another type of change comes in the form of legal action. For example, in the Yelp case, close to 700 lawsuits were filed in attempts to force Yelp to change its review filtering algorithm [285]. Though none of the lawsuits in the Yelp example were successful, they had the potential to put the platform in a position in which it would have been required to respond and enact change or risk behaving illegally.

There are also less direct repercussions of the legal action. Regardless of whether a lawsuit has the desired result for the plaintiff, the filing of a legal case in itself creates publicity for the issue.

Perhaps the most effective type of change is at the platform level. Sometimes when an everyday algorithm audit brings a problem to the forefront, the creators or employers of the algorithm take notice and agree that a change should occur. An official statement can create legitimacy around the audit and the identified issue, and swift redress can occur, since those who have decided on change now include those who have the power to enact change. Both of these resulted from the Twitter racial cropping case, as Twitter posted on the company blog [8] and committed to changing the way that image previews are cropped. In another example, Google revised its image recognition algorithm's racial bias in classifying some Black users as gorillas (which was found and reported by everyday users) [548]. However, platforms' attempts at redress are not always effective. For example, Google preemptively announced in 2020 that it had fixed its translation classifier to address bias that assigned gendered pronouns in an often stereotypical way to professions and activities when translating gender-neutral languages [387, 574]; the issue reemerged in 2021 with other languages (e.g., [177, 330, 536]).

When platforms fail to act to fix a harmful and biased algorithmic behavior, everyday users sometimes go beyond detecting, testing, and reporting a bias to begin rectifying the behavior themselves. For example, in the Booking.com case, users started manipulating the algorithm to return lower algorithmically calculated ratings in their hotel reviews so that the actual ratings better aligned with users' intended ratings [168]. Everyday users also engage in what we call "collective repair" to resist and ameliorate the harm an algorithm can bring to a community. For example, in an everyday audit TikTok users found that the platform's For You Page algorithm suppresses content created by people with certain social identities, such as LGBTQ+ users and people of color. In response, some users worked together to amplify the suppressed videos by specifically engaging with them and adding comments [288, 485].

3.6 Design Implications

We have discussed users' auditing behaviors around the Twitter image cropping algorithm and around rating algorithms on Yelp and Booking.com. Across these cases of everyday algorithm auditing, we observed similar patterns: a group of day-to-day users comes together to detect, understand, and interrogate problematic machine behaviors during the course of their ordinary interactions with algorithmic systems. In this section, we explore what lessons might be drawn from the study of everyday algorithm auditing to overcome limitations of previously proposed auditing approaches. We ask how designers might facilitate everyday algorithm auditing practices, building upon regular users' existing motivations to engage in such auditing behaviors and supporting their unique strengths in surfacing harmful algorithmic behaviors. How might designers support and scaffold everyday algorithm audits, without intervening so heavily as to sacrifice their bottom-up, user-driven nature?

Below we outline five broad categories of potential design interventions—a) community guidance, b) expert guidance, c) algorithmic guidance, d) organizational guidance, and e) incentivization—and discuss potential design trade-offs.

3.6.1 Design for Community Guidance

Across our cases, we observed different degrees and forms of collaboration among everyday auditors. In general, more collective efforts may lead to more impactful everyday algorithm audits. For example, everyday auditors in the Yelp case engaged in more collaborative auditing behaviors compared with the Booking.com case. We could speculate that, through increased interaction with other everyday auditors, Yelp users may have been more likely to form “counterpublics” against the dominant algorithmic system, in a collective effort toward sensemaking and consensus building. Moreover, more collective everyday auditing efforts may be more successful in raising awareness and pushing for changes on the algorithmic, platform, and even societal level, as we observed in our discussions around Twitter’s cropping algorithm.

One way designers might support collective auditing behavior is to offer mechanisms that help everyday auditors guide each other’s efforts. At minimum, this might include providing spaces for community discussion so that everyday auditors can discuss, raise awareness about, and build upon each other’s findings, as we have already observed in the Yelp and Twitter cases. Apart from directly facilitating the auditing process, public discussion forums may help garner support from advocacy groups and news organizations to raise awareness and push for potential remediations on different levels, as we observed in the YouTube LGBTQ+ demonetization case.

Beyond simply providing spaces for discussion, we might imagine tailoring or augmenting discussion channels with mechanisms that are explicitly designed to support everyday auditing. For example, a discussion platform designed for everyday auditing might allow community members to upvote specific algorithmic behaviors that other community members have reported, in order to collectively surface the most *severe* reports, or reports that may *benefit from further discussion*. Furthermore, when an everyday auditor selects a particular report, the discussion platform might provide an overview of relevant hypotheses and findings that other auditors have generated so far, to help the auditor build upon previous contributions instead of retreading old ground.

When designing to support community guidance, we should also consider the possibilities of leveraging existing platforms rather than creating entirely new ones. In the Yelp case, users were able to leverage embedded discussion forums to organize their everyday auditing. In the Twitter case, the platform being audited was itself a discussion platform, allowing users to discuss and test algorithmic behavior *simultaneously*. This evokes Grevet and Gilbert’s “piggyback prototyping” technique in which existing, popular platforms (e.g., Facebook and Twitter) are appropriated in the design of new social computing systems in an attempt to help overcome the challenges of obtaining critical mass on entirely new platforms [222]. A potential trade-off is that hosting community discussion *within* the same platforms that are under audit may lead to concerns about platform-driven censorship or suppression, given uneven power dynamics and the potential for value conflicts between justice-oriented audits and profit-oriented platforms. We anticipate that in certain cases, everyday auditors may feel more comfortable hosting these discussions elsewhere.

3.6.2 Design for Expert Guidance

Although the cases we reviewed were often initiated organically by situated users, others with relevant technical expertise sometimes weighed in as well. In the Twitter

photo cropping case, developers from Twitter intervened in the discussion of racial bias to weigh in with some of the background knowledge they had available on how the cropping algorithm functions [559]. We expect that everyday algorithm audits can benefit from expert input, whether from system developers, researchers, or regulators. However, it remains an open question what forms and degrees of expert engagement may be most helpful in facilitating everyday audits.

To help steer everyday auditors' ongoing efforts in more productive directions, designers might invite relevant experts to continuously monitor community discussions and intervene as needed [107, 158]. For example, at any stage of an everyday audit, a product team might weigh in and provide feedback, based on technical knowledge of their products, regarding the plausibility of user-generated hypotheses. However, if this is not handled carefully, experts' involvement may risk diminishing everyday auditors' sense of community and autonomy. For example, the insertion of technical or domain experts into the middle of an everyday audit might create unequal power dynamics in community discussions.

Relevant experts may also provide guidance to users at the initiation stage. For example, to complement their own internal auditing efforts, a product team may actively draw users' attention to suspected issues or components of a system that they would like users to audit, then suggest testing strategies for users to try [256]. Similarly, in the case of ImageNet Roulette [130], researchers built an interface to help everyday users quickly and easily audit a model trained on the ImageNet dataset, then disseminated this interface on social media. In turn, the ImageNet Roulette interface went viral, spurring a large and diverse group of users to test for algorithmic biases using images of their choosing. Although audits were initiated by technical experts in both of the above examples, users still had agency and autonomy in directing their collective efforts via social media.

The design of bidirectional feedback mechanisms between developers and everyday users of algorithmic systems represents a critical direction for future research. In addition to the need for systems that help developers provide effective feedback and guidance to users who are engaged in everyday auditing behaviors, there are opportunities to design systems that scaffold users in providing more useful, actionable feedback to developers.

3.6.3 Design for Algorithmic Guidance

For certain auditing tasks, such as exploring a system's behavior, gathering evidence, or testing hypotheses, everyday auditors may benefit from algorithmic assistance. For instance, designers might develop auditing interfaces that automatically surface potentially important instances for everyday auditors to examine further. As a foundation for such interfaces, it may be possible to build upon emerging algorithmic techniques for crowd-in-the-loop detection of "unknown unknowns" in ML models (e.g., [48, 98, 314, 332, 378, 535]), which automatically surface cases that are more likely to be mislabelled and/or misclassified. These methods focus on surfacing regions of a model's error space in which the model is highly confident yet incorrect [314]. This form of algorithmic guidance could supplement guidance that users may receive from each other (e.g., via upvoting mechanisms) and from expert guidance (e.g., suggestions for how everyday auditors should direct their search efforts), potentially surfacing cases that these other forms of guidance would miss.

In some of the everyday auditing cases we have reviewed in this chapter, establishing the presence of certain issues requires some form of *comparison*. We distinguish between two major categories of auditing behaviors: "instance-based" and

“comparison-based” auditing. Instance-based auditing applies to cases where the observation of an isolated harmful algorithmic behavior is sufficient to ground an argument for remediation. For example, when users found that the Google Photos app assigned a photo of Black people with the label “gorillas” [231], it was not necessary to demonstrate that most photos of Black people receive the same label, nor was it necessary to demonstrate that Black people are consistently assigned this label at a higher rate than white people. Rather, users felt that the app should *never* assign such a harmful label to images of Black people. By contrast, comparison-based auditing applies to cases where, in order to establish that a harmful bias truly exists and is worth addressing, it is necessary to compare multiple instances and demonstrate the existence of statistical patterns. For example, when Sweeney [512] found that a web search for her own name yielded arrest record ads, she wanted to understand whether systematic discrimination was present. To do so, it was necessary to search for many Black sounding names and statistically compare the results against those for other names [512]. To support such forms of comparative auditing, designers might provide interfaces and algorithmic guidance to empower everyday auditors, who typically lack knowledge of statistics or causal modeling, in testing statistical or causal hypotheses.

3.6.4 Design for Organizational Guidance

In our cases, we also observed fluid divisions of labor among users across the lifetime of an everyday algorithm audit. For example, in the Twitter cropping case, some users largely focused on tweeting images to test how the image cropping algorithm would behave for each. Meanwhile, other users contributed more in retweeting and raising awareness about observed issues, or in forming hypotheses about observed algorithmic behaviors.

Designers could help guide the division of labor through interventions that encourage everyday auditors to take on particular roles during an auditing process, including roles that are expected to be important and might not arise organically. For example, across all of our cases, a major challenge was a lack of effective synthesis mechanisms to help users keep track of each other’s findings and hypotheses and build upon each other’s contributions so far (cf. [128]). “Synthesizer” may be a valuable role for some everyday auditors to take on, in order to help the collective make more cumulative progress. More broadly, we could imagine organizing everyday auditors along a number of major roles, such as (1) *Initiators*, who focusing on identifying problematic algorithmic behaviors; (2) *Generalizers*, who examine instances flagged by initiators, and use these to inform hypotheses about the scope of the issue and possible underlying causes; (3) *Amplifiers*, who help share and broadcast what has been discovered to others, raising awareness of findings, hypotheses, and relevant discussions so far; and (4) *Synthesizers*, who transform the outputs of this iterative auditing process into summaries for others to build on.

3.6.5 Design for Incentivization

Finally, while community, expert, algorithmic, and organizational guidance may help to improve ongoing auditing processes, another major design opportunity is to explore mechanisms that incentivize everyday auditors to have more active and sustained participation. For example, one potential design direction is to offer an “algorithmic bias bounty program,” drawing inspiration from “bug bounty” programs in cybersecurity. The core idea of security’s “bug bounty” program is to incentivize

hackers to share vulnerabilities with legitimate organizations for monetary and reputational rewards, as an alternative to selling or exploiting those vulnerabilities [392, 452]. Similarly, in the context of everyday algorithm audits, we could encourage users to directly report detected biases to organizations and platforms, offering social incentives (e.g., reputations and points) and monetary incentives accordingly. A potential risk is that, if not designed carefully, the provision of extrinsic rewards for algorithm auditing may reduce useful spontaneity or may inadvertently diminish users' existing intrinsic motivations to engage in everyday auditing behaviors.

Indeed, while this chapter was under review, Twitter announced its own "bias bounty" program, inspired by cybersecurity "bug bounty" programs and by its own image cropping cases. This program aimed at incentivizing people to detect, test, and report potential biases [114]. However, such "bias bounty" programs also heavily rely on participants with relevant technical expertise to detect biases, similar to bug bounty programs in cybersecurity, which rely on experienced hackers to identify bugs. We see design opportunities here to develop systems that can combine strengths of both technical experts and everyday users. For example, an audit could begin by inviting a wide and open range of everyday users to participate in surfacing harmful algorithmic behaviors. Initial hypotheses and observations from everyday users could then serve to inform further, systematic investigations by machine learning experts and analysts. The initial, open phase of auditing may uncover issues that a narrower group of technical experts might otherwise have missed. Indeed, in some ways, this is how the Twitter "bias bounty" program proceeded, given that participants in the program were directly informed by observations and hypotheses that everyday users had developed previously.

In addition to incentivizing auditors, there is also a need to incentivize developers and platforms both to implement these measures (e.g., supporting user-led audits on embedded discussion forums) and to remediate issues that users uncover. Holstein et al. (2019) found that commercial product teams, fearful of negative public attention, are often motivated to uncover harmful algorithmic behaviors as early in the process as possible. Acknowledging the limitations of existing auditing approaches, several industry practitioners interviewed in this research expressed interest in implementing mechanisms for user-led auditing [256]. Some product teams were motivated enough to experiment with developing their own collective auditing tools and processes. While these findings highlight that some product teams and companies are already motivated to implement such measures, others may be hesitant to encourage users to scrutinize their systems. Prior work has demonstrated that, even where motivation is lacking, public awareness raising around harmful behaviors in commercial AI systems can motivate target companies to prioritize remediation [414]. Similarly, recent work from Vincent et al. [552] discusses potential ways for the public to withhold their data contributions as leverage, reducing the effectiveness of specific data-driven technologies in order to motivate companies to address users' concerns. Finally, given the uneven power dynamics between users and platforms, structural interventions are also needed (e.g., policy and legal interventions) to ensure external accountability.

3.7 Discussion

We have taken initial steps to theorize and explore an under-studied phenomenon in which everyday users detect, interpret, question, and bring attention to problematic machine behaviors in their day-to-day interactions with algorithms. We argue

that such “everyday algorithm auditing” is especially powerful in detecting harmful behaviors that are challenging for existing auditing approaches discussed in the academic literature. In this section, we discuss the unique power and contributions of everyday algorithm auditing and raise a few open questions on how to better support it for future research.

3.7.1 The Power of Everyday Algorithm Auditing

Past research has highlighted the limitations and blindspots of existing auditing approaches and speculated that everyday users might be able to overcome these limitations (e.g., [169, 256]). In this chapter, we take a step further by studying how such user-led audits unfold in the real world and by exploring how we might better support these processes. Through detailed case studies, we have shown that everyday users are able to detect harmful algorithmic behaviors that are challenging for traditional auditing approaches and that users can exercise agency and autonomy in directing their own auditing process. Below we highlight a number of unique contributions of everyday algorithm auditing and how it can help overcome some of the limitations presented in existing approaches.

First, everyday algorithm auditing leverages the *lived experiences* of everyday users. Past literature suggests that existing approaches often fail to involve auditors with relevant cultural backgrounds and lived experience that are critical to detect sensitive harms [589]. Everyday algorithm auditing, with contextually situated users, is especially powerful to appropriately identify the problems at hand. Although a crowdsourced/collaborative auditing approach [447] also invites users’ participation, it often relies on crowdworkers, who do not necessarily have lived experiences with the algorithms being audited, do not communicate and work together toward a common goal, and therefore might not be sensitive to the related harms.

Second, everyday algorithm auditing engages and leverages the *situated knowledge* of everyday users. As past literature suggests, many harmful machine behaviors are challenging to detect outside of situated contexts of use [127, 190, 256, 343, 462]. Through their day-to-day interactions with an algorithmic system, everyday users are particularly well positioned to detect these types of behaviors that emerge in real-world contexts of use, in the presence of complex social dynamics, and in the changing norms and practices of using algorithmic systems over time.

Third, in an everyday algorithm audit, users are able to form *counterpublics* via different social media channels or community forums and participate in their own collective sensemaking and consensus building. This is especially powerful since only in such collective attempts, will users be able to build on each others’ contributions, test each others’ hypotheses, and support each other in different forms (e.g., helping publicize their efforts). In contrast, in a crowdsourced/collaborative audit [447], users are often working on individualized tasks that have been assigned to them, without participating in discussion.

Finally, in an everyday algorithm audit, users have *control* over the auditing process. This differs from a crowdsourced/collaborative audit [447], in which users are hired via crowdsourcing platforms to work on decomposed subtasks determined entirely by outside parties; in an everyday algorithm auditing, users—often collectively—decide their own course of action. Their autonomy and agency help steer the audit in the directions that are most useful and meaningful for them, which are often overlooked by existing approaches.

3.7.2 Appropriate and Timely Interventions

We have discussed five types of design interventions we might offer to support everyday algorithm auditing: community guidance, algorithmic guidance, expert guidance, organizational guidance, and incentivization. We have also discussed some potential trade-offs that we need to consider when offering these types of interventions. Below we raise open questions regarding the appropriate timing and proper degree of intervention for future research.

When to intervene and when to stop

We have discussed that interventions can happen at any stage of an everyday algorithm audit and that different stages might benefit from different types of interventions. For example, we might anticipate that having expert guidance during the initiation stage (e.g., as in the case of ImageNet Roulette) and the hypothesizing and testing stage (e.g., in the Twitter racial cropping case) might be especially helpful. On the other hand, community guidance might be more useful at the awareness raising stage (e.g., as in the YouTube LGBTQ+ demonetization case) and the hypothesizing and testing stage (e.g., in the Yelp and Twitter cases). However, currently we have very limited understanding of the appropriate timing for intervention. For example, how can we maintain the organic nature of everyday algorithm auditing, especially at the initiation stage, but also prompt a group of people to start auditing first, which might inevitably remove some of the organic nature?

How much intervention

We have also discussed different forms of intervention and some hypothesized trade-offs. For example, introducing expert guidance might create unequal power dynamics in community discussions and risk diminishing users' sense of community and autonomy. Hosting community discussions within the same platforms that are under audit may lead to concerns about platform-driven censorship or suppression. Currently we are only at the starting point to understand and explore the appropriate degrees of these interventions. How much intervention is too much? What are the trade-offs in moving from more organic to more organized at different phases of an everyday audit? These questions need further research and investigation; the exploration of everyday algorithm auditing is in its infancy, and our work presents a first step.

3.8 Conclusions

In this chapter, we have drawn on real-world case studies and prior theories of everyday resistance and counterpublics to understand how everyday users—either individually or collectively—detect, understand, and interrogate problematic machine behaviors in their day-to-day interactions with algorithmic systems. By comparing the lifetime and dynamics of several cases of everyday algorithm auditing, we have proposed a process-oriented view of everyday algorithm audits, involving initiation, awareness raising, testing and hypothesizing, and remediation. We have drawn lessons from these cases for future research and design around everyday algorithm auditing, outlining five broad categories of potential design interventions to support everyday audits—a) community guidance, b) expert guidance, c) algorithmic

guidance, d) organizational guidance, and e) incentivization—along with potential trade-offs.

As the first work conceptualizing and studying everyday algorithm auditing, this research is highly exploratory in nature. The lifetime and dynamics presented as part of an everyday audit are intended to be representations of what we see at this stage of exploration, capturing both current practices we have observed and possibilities we see for future practice. As such, these represent a set of untested assumptions about everyday algorithm audits that could change over time as more empirical work is conducted in this space. That is, neither the cases presented in this chapter nor the approaches we used to analyze them are comprehensive: they outline an observed phenomenon and serve as illustrative examples to better understand how everyday audits work, what paths everyday audits could take, and what impacts everyday audits could have. Future research should conduct in-depth follow-up studies with users to investigate the practices and strategies they use to find and make sense of harmful algorithmic behaviors. Another fruitful, complementary area for future research includes an exploration of the design opportunities discussed above. For example, how can we design platforms that support users in conducting more effective everyday algorithm audits, both individually and collectively?

In sum, this work is an initial step toward bridging the gap between algorithm auditing approaches in academia/industry and everyday auditing behaviors that emerge in day-to-day use of algorithmic systems. It is my hope that this work will help to inform future research on everyday algorithm auditing, as well as the design of future platforms and tools that empower people to meaningfully audit the algorithmic systems that impact their lives every day.

Chapter 4

Investigating Users' Strategies for Uncovering Harmful Algorithmic Behavior

This chapter is adapted from my paper [148]:

Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, Motahhare Es-lami. "Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior." In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. (2022).

Recent work in HCI suggests that users can be powerful in surfacing harmful algorithmic behaviors that formal auditing approaches fail to detect. However, it is not well understood how users are often able to be so effective, nor how we might support more effective user-driven auditing. To investigate, I conducted a series of think-aloud interviews, diary studies, and workshops, exploring how users find and make sense of harmful behaviors in algorithmic systems, both individually and collectively. Based on the findings, I present a process model capturing the dynamics of and influences on users' search and sensemaking behaviors. I find that 1) users' search strategies and interpretations are heavily guided by their personal experiences with and exposures to societal bias; and 2) collective sensemaking amongst multiple users is invaluable in user-driven algorithm audits. I offer directions for the design of future methods and tools that can better support user-driven auditing.

4.1 Introduction

The presence of biases and inequities in algorithmic systems has led researchers to develop new approaches for algorithm auditing to detect biased, discriminatory, or otherwise harmful behaviors¹ (e.g., [91, 127, 168, 256, 323, 379, 417, 447, 512, 597]). Typically these auditing techniques are led by experts such as researchers, activists, industry practitioners, and government agencies [106]. For example, in "scraping audit" techniques, experts query an algorithmic system and investigate and compare the outputs [447]. As another example, in a "sock puppet audit", experts use computer programs to impersonate different types of users, inject fake data into the system, and evaluate the results [447]. Although expert-led auditing approaches

¹When relevant, we differentiate between harmful biases and harmful behaviors more broadly in algorithmic systems. While these categories often overlap, not all algorithmic harms are best understood as "bias," and not all algorithmic biases are necessarily harmful [74].

have been greatly impactful, the absence of the actual context of use and everyday users in the auditing process can still result in major blindspots in practice. As well as experts' cultural blindspots, social dynamics, changing norms, and new circumstances can hinder experts' detection of many types of algorithmic biases and harms [127, 190, 254, 462, 474, 589]. In contrast, recent years have seen many cases in which *users* uncover and raise awareness about harmful algorithmic behaviors that they encounter while interacting with algorithmic systems [168, 169, 249, 474]. Some examples include Twitter users detecting racial bias in Twitter's image cropping algorithm [249], Apple Card holders uncovering gender bias in Apple's credit limit algorithm [546], and a group of YouTubers reporting the demonetization of LGBTQ+ content by the YouTube recommendation algorithm [433].

Inspired by these observations, researchers are beginning to explore ways to support the power of users and crowds in order to overcome limitations of expert-led algorithm auditing approaches [474]. For instance, researchers evaluated personalization impacts of search algorithms as well as price steering algorithms on e-commerce websites by crowdsourcing the outputs of these algorithms and evaluating the results [240, 241]. Another related line of research has developed prototype interfaces and crowdsourcing pipelines intended to support people in actively searching for algorithmic errors, biases, or harmful behaviors (e.g., [43, 98, 130, 383, 505]).

While these efforts have demonstrated promising results, it remains poorly understood *how* people go about searching for and making sense of harmful algorithmic behavior, let alone how they are often able to be so effective at doing so. In order to guide the design of more effective supports for user-driven algorithm auditing processes, it is critical to understand how these processes unfold, what search and sensemaking strategies users are likely to use, and where these strategies may be likely to break down. To begin to address these issues, in this chapter, we explore the question: **How do people find, make sense of, and evaluate potentially harmful algorithmic behavior?**

We conducted a three-phase study with 23 everyday users of algorithmic systems to investigate how they surface and make sense of potentially harmful algorithmic behaviors. In the first phase of this study, we conducted think-aloud interview sessions in which individual participants were tasked with actively searching for and making sense of potentially harmful algorithmic behaviors in the context of Google Images search engine. To complement these live observations, participants then took part in a 14-day diary study in which they were asked to report potentially harmful algorithmic behaviors that they found both through active searching and incidentally during the course of their day-to-day interactions with algorithmic systems. Finally, we invited participants to participate in group workshop sessions and work collectively to interpret and discuss some of the cases that were uncovered during the diary study.

Throughout our study, participants showed great ability to detect and reason about potentially harmful algorithmic biases. For example, participants submitted a total of 160 evidence-supported reports of potential biases or other harmful algorithmic behaviors in the diary study, spanning gender, sexual orientation, race/ethnicity, socioeconomic class, age, body type, disability, religion, politics, and combinations of these. Our analyses yielded a three-stage model capturing the process by which people uncover biased and harmful algorithmic behaviors: 1) *search inspiration* encompasses ways in which participants came up with ideas for where to look for potential biases in an algorithmic system, 2) *sensemaking* involves the ways that participants understood and evaluated algorithmic behaviors that might be harmfully biased, and 3) *remediation* covers participant actions and desires to mitigate the

ramifications of harmful algorithmic bias. Connected to each stage of this process, we uncovered two common influences on the process across participants: knowledge and beliefs, and platform affordances. Participants used their prior knowledge and beliefs related to biased and harmful algorithmic behavior (i.e., their conception of bias and harm, exposure and experience, expectations and values, and folk theories) to make sense of problematic algorithmic behaviors they encountered. Participants' abilities to detect and make sense of such behaviors were enabled and guided by platform affordances.

Based on our findings, we offer directions for the design of future methods and tools that can better support user-driven algorithm auditing. We discuss benefits, limitations, and challenges of user-driven algorithm auditing, informed by our process model. In doing so, we look to ways that future user-driven algorithm auditing platforms and tools might be designed to best leverage the benefits, mitigate the limitations, and navigate the tensions.

In summary, this work contributes the following to the HCI and CSCW communities:

- The first in-depth empirical investigation of how users find and make sense of harmful behaviors in algorithmic systems, both individually and collectively
- A conceptual process model capturing the ways that users find and make sense of biased and harmful algorithmic behaviors, developed through an analysis of our study participants' actions and explanations
- Directions for the design of future methods, platforms, and tools to support user-driven algorithm auditing, informed by our process model

4.2 Related Work

4.2.1 User-Driven Algorithmic Auditing

Recent years have seen many real-world cases in which *users* uncover and raise awareness about harmful algorithmic behaviors that they encounter through their day-to-day interactions with algorithmic systems, both individually and collectively [168, 249, 474]. In a recent, highly publicized example, Twitter users noticed problematic behavior from Twitter's image cropping algorithm and suspected that the algorithm was prioritizing white faces over Black ones (e.g., [86]). Twitter users came together organically to investigate. They tested Twitter's cropping of various images to assess whether the cropping algorithm exhibited a systematic, harmful bias against Black faces. They built upon each other's findings in online discussions to present evidence or counter-evidence for patterns discovered by other users or present new evidence and hypotheses of their own. Similar cases of "everyday algorithm auditing" have been observed around a range of algorithmic systems, including image search engines, online rating/review systems, machine translation, personalized advertising, and image captioning [474].

Researchers have begun to explore the design of systems that can support the power of everyday users and crowds to surface harmful algorithmic behaviors that formal auditing approaches might otherwise miss [474]. A related line of work has developed prototype interfaces and crowdsourcing pipelines intended to support people in actively searching for algorithmic errors, biases, or harmful behaviors. For example, ImageNet Roulette was a simple online interface developed by artists and researchers to support users in exploring and interrogating the input/output

space of an image captioning model trained on the ImageNet dataset. This project provoked discussions on social media, as users shared findings and hypotheses, and sometimes built upon each other's observations [130, 363]. In another project, Ochigame and Ye developed a web-based tool called Search Atlas, which allows users to explore and easily compare Google search results as if they were located in different countries [383]. Another body of work has explored the design of crowd pipelines, interactive visualizations, and interfaces to support crowdworkers in searching for and making sense of algorithmic errors (e.g., [43, 98, 505]). For example, Attenberg and Ipeirotis developed a game-like task called "Beat the Machine," which asks crowdworkers to find cases where an automatic predictive model is incorrect. While these efforts have demonstrated promising results, many open questions remain regarding how everyday users go about uncovering harmful algorithmic behavior, let alone how they are often able to be so effective at doing so.

4.2.2 Information Search & Sensemaking

In this chapter we build upon existing work in the realm of information search and sensemaking to understand how people go about surfacing harmful algorithmic behaviors both individually and collectively. Pirolli and Card [400] proposed a general model of information search and sensemaking, based on cognitive task analyses conducted with data analysts. Their model posited four major phases of the sensemaking process, which may proceed through iterative cycles rather than in a linear sequence. These phases included information gathering, the representation of information in ways that aid analysis, the development of insights through manipulation of this representation, and the creation of some knowledge product or direct action based on these insights. Pirolli and Card's model organizes these phases into two major loops: the foraging loop and the sensemaking loop. At a high-level, these loops map onto the phenomena we investigate in this chapter: how users search for and make sense of harmful algorithmic behaviors.

In the sensemaking literature, both search and sensemaking are understood to be *influenced by* the prior knowledge and beliefs of the sensemaker(s), as well as to *influence* their knowledge and beliefs [87, 281, 308, 439]. As such, we might expect that the strategies people use to search for and make sense of harmful algorithmic behaviors will be shaped, in part, by folk theories they hold about the workings of a given algorithmic system [145, 165] and by their existing perceptions and beliefs about algorithmic bias and unfairness in a given context [62, 66, 88, 136, 223, 577] — both of which are active areas of research in their own rights. However, it remains unknown how exactly such prior knowledge and beliefs inform people's search and sensemaking strategies in the context of everyday algorithm auditing, and it is unclear what other individual or group-level factors might influence these search and sensemaking processes. The current research aims to build on this prior research as a foundation to explore these questions, with the ultimate goal of informing novel design interventions to support more effective everyday algorithm audits in practice.

4.3 Methods

In this study, we sought to investigate how people uncover harmful behaviors in everyday algorithmic systems: How do people find, make sense of, and evaluate potentially harmful algorithmic behavior? To explore this question, we conducted a

three-phase study with 23 participants. We first conducted think-aloud interviews in which participants were tasked with actively searching for potentially harmful algorithmic behaviors online and making sense of these behaviors. These think-aloud interviews provided an opportunity to closely observe participants' thought processes and ask clarifying questions in a relatively controlled setting. Second, to complement these live observations, participants then took part in a 14-day diary study in which they documented harmful algorithmic behaviors that they found through active searching or during the course of their day-to-day interactions with algorithmic systems. The diary study enabled observation of participants' search and sensemaking strategies in a more naturalistic setting and over a longer period of time, allowing for the possibility of chance encounters with harmful algorithmic behaviors. Third, we then invited participants who completed the think-aloud interview and diary study to participate in group workshop sessions. In these sessions, participants were tasked with working together to evaluate and discuss some of the cases that were uncovered during the preceding diary study. We describe these phases in more detail in the following sections.

4.3.1 Recruitment

We recruited participants via social media (i.e., Twitter, Facebook groups, and Nextdoor), Craigslist Jobs, and mailing lists for local community organizations. Participant diversity was particularly important in this study as previous work suggests that lack of diversity in auditing teams may result in blindspots in finding harmful algorithmic behaviors [256, 474, 589]. Therefore, the recruitment blurb directed those interested to a demographic information questionnaire to provide characteristics including gender, race/ethnicity, sexual orientation, religion, political views, employment status, income, locations lived, languages spoken, level of education, tech savviness, and familiarity with algorithms. We pulled these characteristics from types of algorithmic bias in previous literature and framed relevant questions by drawing from a combination of other HCI research [457], the US Census survey [93], and a conference panel including Pew Research Center advice on categorizing religious groups [257]. Sixty-five people completed this screening form, from which we invited people spanning a diverse range of identity characteristics and backgrounds to participate. We first focused on inviting people from marginalized identity groups, as they tend to be underrepresented in research, then invited people with characteristics we lacked in our already signed up group. For example, after the first set of invites, we noticed a gender skew toward women in our signups, so we began to focus on recruiting men and non-binary people, inviting any who completed the questionnaire.

4.3.2 Participants

The screening process yielded a group of 23 participants (61% women, 30% men, 9% nonbinary) aged 35.2 on average with an age range from 20 to 64. Thirty-five percent of the participants were white, 30% Asian, 26% Black, 9% Hispanic or Latino, 9% American Indian or Alaska Native, and 4% Middle Eastern. (Participants could select multiple races/ethnicities.) See the Appendix A for more participant demographic details. Despite research recruitment commonly lacking enough people from marginalized identity groups, we noticed that we had somewhat less difficulty recruiting participants from marginalized groups (e.g. Black, LGBTQ) than expected. Additionally, no one who expressed interest in participating self-identified

as “very conservative” politically, though that was a presented option, and we recruited half as many men as women participants despite inviting every man who filled out the demographic survey to participate. We speculate that these trends might arise from the topic’s relevance to marginalized groups (and, conversely, diminished personal relevance to those who are not marginalized), mirroring recruitment patterns observed in prior studies related to algorithmic fairness and bias (e.g., [256, 343]).

The think-aloud interviews and diary study took place from November to December 2020, and the workshops occurred in February 2021. Participants were compensated with \$20 for the think-aloud interview. In the diary study, participants received \$5 per submission (up to a maximum of \$30). Finally, participants received an additional \$25 for participation in the workshop. We refer to the participants as P1–P23 throughout this chapter. Unless mentioned otherwise, quotes are taken from think-aloud interviews.

4.3.3 Bias & Harm

As a study about bias and harm, our definitions and participants’ understandings of these concepts are important in interpreting our results. In this study, we first asked participants about their understandings of these terms. This exercise revealed that participants did indeed tend to view bias and harm as distinct, but related concepts. Participants were then told that the current study was focused on “*possible cases of harmful bias and discrimination in algorithmic systems*”, which in turn was defined broadly in terms of instances where an algorithmic system “*behaves in ways that might cause harm for certain individuals, groups of people, or society.*” We purposefully used this expansive definition to ensure that participants had enough of an understanding to complete the study tasks, while at the same time trying to avoid pushing participants toward overly restricted definitions. In the interest of investigating participants’ perceptions, we wanted to avoid prescribing precisely what it means to be more or less harmful or biased.

4.3.4 Phase 1 – Think-Aloud Interview: Individual bias search & sense-making

We conducted 23 remote think-aloud interview sessions, each approximately one-hour long, over Zoom. We chose Google Images as the domain for the think-aloud portions of our interview, given both that it was likely to be familiar to participants and because many historical cases of users finding harmful algorithmic behaviors have taken place in the context of image search [203, 216, 227, 294, 461]. For example, in a widely publicized case, users discovered that Google Images search results for the term “unprofessional hairstyles” overwhelmingly consisted of photos of Black women, compared with search results for the term “professional hairstyles” [18, 268].

The interview consisted of two steps. First, we asked participants to search for specific images. These image search tasks were chosen based on previous work that showed or research team discussions that suggested the image search results might be discriminatory or negatively impact some users. This step aimed to understand users’ perceptions and evaluations of algorithmic behaviors anticipated to be biased in a harmful way. In the second step, participants started looking for new cases of algorithmic bias by exploring, generating, and trying out new ideas for search queries. This step aimed to aid our understanding of how users search for and make

sense of potentially biased algorithmic behavior. Below we describe these steps in more detail.

Evaluating Existing Biases

At the beginning of the interview, we briefly described harmful bias and discrimination in algorithmic systems, then asked participants whether they had heard of this and, if so, what they thought it meant as well as any thoughts and reactions about times they might have experienced it in their own lives. We then asked participants to search for concepts that were known to be associated with societal biases across gender, race, sexual orientation, body image, religion, and culture, as based on prior work and research team investigations. We intentionally selected a diverse sample of tasks to try to avoid biasing participants toward certain kinds of biases and harms. This list of tasks included asking participants how they might find an image of a) a librarian, as the results showed mostly white women, b) a thug, as the results showed mostly Black men, c) one person nagging another, as the results showed mostly heterosexual couples where a woman nagged a man, d) a beautiful woman, as the results showed mostly thin, white women, and e) a wedding, as the results showed mostly Western culture weddings based in Christianity. We acknowledge that our perceptions of bias within these tasks may not represent the only biases present in them (indeed, participants identified others we missed); we used these categorizations merely as a method to select cases to use in the tasks. The cases appeared in the form of targeted think-aloud tasks that asked participants to search for images (using their choice of keywords), then interpret and discuss the search results.

Our motivation for starting with these targeted tasks was two-fold. First, we wanted to see whether and how participants detected biases in the context of search queries that are known to have problematic results. Second, we used these initial, targeted tasks to get participants used to finding and making sense of harmful biases in the context of Google Images, in preparation for the less scaffolded, more open-ended tasks to come. After a participant completed a given targeted search task and interpreted the search results on their own, we probed 1) whether participants thought the search results were biased or discriminatory in harmful ways that might negatively impact people, 2) why (and how) or why not, and 3) what they would do if they had the option to change/adjust the search results.

Looking for New Biases

In the next portion of the interview, we wanted to learn how people go about actively seeking out new cases of harmful algorithmic behavior and what factors influence where and how they look (e.g., what search queries they decide to try, in the context of image search). Thus, participants were instructed to perform the more open-ended task of finding new examples of image search queries that might yield problematic image search results. As participants explored, generating and trying out new ideas for search queries, they were asked to think aloud to provide insight into their thought processes. Interviewers asked follow-up questions in order to better understand how participants generated ideas for search queries to try, why they searched for the specific terms they did, and how they interpreted the image search results they saw in response. At the end, we asked participants what they might do if they encountered something online like what they did in the session's tasks, what

they thought should happen when they do an image search online, and for general session thoughts.

4.3.5 Phase 2 – Diary Study: Finding harmful algorithmic behavior in everyday life

We next invited interview participants to take part in a 14-day-long diary study in which participants were asked to look for potential cases of harmful algorithmic bias or discrimination online. The goal was to better understand how people encounter and make sense of harmful algorithmic behaviors on their own, either through their everyday interactions with algorithmic systems or by actively searching for such behaviors. Participants were asked to record observations of potentially harmful algorithmic behavior via a screenshot; a brief textual description; and an explanation of why they thought it might be harmful or problematic, whom it might harm, and how they found this case. 22 interview participants submitted reports, submitting just over 7 reports on average with the 4 as the lowest number of reports submitted by an individual and 15 as the highest. We encouraged participants to submit cases of harmful algorithmic behavior they weren't entirely certain about, giving them the freedom to explore their thoughts and suspicions in different situations without needing validation.

Interview & Diary Study Data Analysis

We conducted a thematic analysis approach of the data from both the think-aloud interviews and diary studies. The two interviewers (the first and second authors of this chapter) performed an initial round of thematic coding by coding portions of the interviews and diary studies separately using NVivo. The goal was to capture a diverse and open range of qualitative observations, including reflections participants shared about how they conceptualize and understand bias and unfairness, the strategies they used to come up with search queries to find harmful algorithmic behaviors, and their perspectives on ways to remediate harmful algorithmic behaviors that they identified. Next, the research team collectively engaged in an iterative process of discussing and clustering codes into higher level themes. Below, we describe these themes.

4.3.6 Phase 3 – Workshops: Collective evaluation & validation of reported biases

Whether they arise incidentally as in the Twitter photo cropping case discussed earlier [249], or are initiated intentionally as in the case of ImageNet Roulette [130, 363], user-driven audits typically involve users coming together to evaluate and validate harmful algorithmic behaviors or potential biases that they or other users have observed [474]. Therefore, after investigating individual users' strategies to uncover harmful algorithmic behavior in the think-aloud interviews and diary studies, we engaged participants in collective follow-up activities. We conducted a set of four 60–90-minute workshops with 3–5 participants each, with the dual goals of a) understanding how people work together to evaluate user reports of potentially harmful algorithmic behaviors and b) beginning to understand how we might design to facilitate such collective activities.

In these workshops, participants were presented with de-identified reports of potentially harmful algorithmic behavior, generated and submitted by individual

participants during the diary study. We curated this set of individually reported biases, drawing from the full set of diary study responses, in order to span a range of “types” of reported biases and harms. Specifically, we conducted a bottom-up tagging of cases submitted in the individual reports, with tags including gender, sexual orientation, race/ethnicity, socioeconomic/professional, age, body depiction, religion, and political focus in the reported issues. Given that our goal was to understand how users work together to evaluate potentially harmful algorithmic behaviors, we selected all submissions in which 1) the submission describes the issue as having to do with algorithms and 2) the submission describes the issue as harmful. This produced a set of 77 user reports. From this set, we randomly selected one submission with each tag (as submissions could have more than one tag, some overlap is present in this set). We also included a few user reports that did not fit the criteria above, motivated by knowing that such noisy reports do appear to arise in practice and wanting to see whether and how workshop participants might differentiate between these and reports of genuine algorithmic behaviors during their discussions. To get these, we randomly selected 3 submissions that did not follow one or both of the conditions above (i.e., one was not algorithmic, one was not harmful, and one was neither). Our entire selection process yielded 11 individual reports for use in the first workshop. After the first workshop, we chose a smaller set of 6 reports because reviewing this set of 11 reports took a substantial amount of time from participants in the first workshop. Since reports could have multiple tags, we removed reports such that we had one instance of each tag in this set of 6 reports.

We asked workshop participants to evaluate these individual user reports and work together to create a collective report that provides insights for someone who might be able to act on them. Each workshop proceeded in two phases. The first phase involved selecting a case to collectively write a report about. To do this, participants individually examined the user reports and advocated for what they thought was the highest priority to create a report for (e.g., in two workshops we asked participants to choose two of the submitted cases to focus on in the collective report and explain their rationales). Based on this, one user report was selected (e.g., based on which were most advocated for by participants) as the focus for the next phase of the workshop. Table 4.1 shows the cases selected in this phase and discussed and written about in the next phase of the workshops.

In the second phase, participants worked together to make sense of the selected user report, asking and answering questions collectively about the reported issue. We provided a report template, designed to structure and prompt workshop participants’ discussions. We generated the first, bare-bones version of this template based on our knowledge of the minimal information that might be important for decision-makers receiving such a report to know about an issue when attempting to address it. This template included prompts to get participants thinking about the desired report recipient, what happened, evidence available and evidence still needed, and actions moving forward. We iteratively honed the template over the course of the workshops based on our observations of how participants interpreted (or misinterpreted) each prompt. To help inform our broader project—toward designing ways to facilitate more effective everyday algorithm audits—we also iterated on other aspects of the workshop implementation across the four workshop sessions. For example, after noticing that participants took on some roles naturally and desired others that did not come as naturally, we began assigning roles in the second phase of the workshop. However, in this chapter, we focus primarily on common themes that we observed across all four workshops.

Workshop Cases	Descriptions
Disability on Google Images	The participant searched “person with disability” on Google Images and the results were primarily about “visible disability” such as the use of canes and wheelchairs.
Pierced earrings on Google Images	The participant searched “pretty ear piercings” on Google Images and the results were all images of white women.
Game developers on YouTube	The participant searched “game design for beginners” on YouTube and the recommendation algorithms kept recommending related videos that target/represent men game developers.
Engineers on Google Images	The participant searched “engineers” on Google Images and the search results were predominantly about “civil engineers.”

TABLE 4.1: Detailed descriptions of the four primary cases (user reports from our diary study) selected by workshop participants.

Workshop Data Analysis

To analyze data from these workshops, our research team conducted a series of four Interpretation Sessions. In these, the research team first went over one workshop’s data at a time to mark important observations as a group. Then, the research team iterated upon these observations through group discussions to cluster them into higher-level categories. Through our interpretation sessions, high-level themes emerged that were grouped with themes from the interviews and diary studies. We describe these themes below.

4.4 Results

Our study results highlight the significant power that users have to surface potential biases and harms in the algorithmic systems they interact with. In the following sections, we explore this as well as a process model developed based on our analysis of the actions our study participants took and their explanations of these actions.

4.4.1 Users’ Power in Detecting Anticipated & Potential Biases & Harms

Participants showed great ability to detect and reason about potentially harmful algorithmic bias in both specific tasks where the research team anticipated bias might be found and in less prescriptive searches where participants directed their own investigations. In their evaluations, participants identified a wide variety of biases. Types of algorithmic biases named by participants included age bias, appearance bias, disability bias, gender bias, job type bias, location bias, cultural bias, political bias, racial bias, skin tone bias, sexual orientation bias, religious bias, and socioeconomic bias.

In terms of the ways that algorithmic behaviors could harm, participants described representational harm (i.e., lack of representation, overrepresentation, misrepresentation), the generation of negative emotions (e.g., offense, anger, annoyance, upsettedness), the perpetuation of incorrect ideas (e.g., misleading stereotypes), and sometimes no harm. Participants considered harm in a variety of ways: they mentioned specific communities or groups of people who would be negatively affected, such as marginalized identity groups and populations vulnerable to being misled (e.g., children and older adults were frequently mentioned); they connected to the potential for concrete physical or allocative repercussions; and they frequently pointed out that harmful algorithmic behaviors impact everyone, not just the most

How this case was found (e.g., Where did you see this? What were you doing? How did you find/notice this? What is the context?):

“I used Google search for images of romantic couples. There images that came up were diverse, they appeared to represent couples from different cultures.”

Why this case might be harmful/problematic & whom it might harm:

“The only image for a biracial couple was a black and white hand holding no actual person, no images of middle aged or elderly couple, no same sex couple, except for a Hallmark movie preview of an upcoming movie featuring 2 males that would be highlighted in the movie. Middle aged, elderly couples of all races and cultures, same sex couples of all races and cultures could be harmed.”

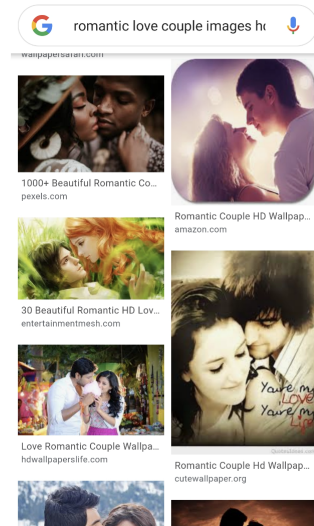


FIGURE 4.1: One example case submitted by our diary study participant P3 (“romantic couples”), including a screenshot of the potential bias, how this case was found, why this case might be harmful and whom it might harm.

directly connected people, through the perpetuation of harmful ideas that damage society at large.

Existing Biases

In both the think-aloud interviews and the diary study, participants detected and made sense of a wide range of potentially harmful algorithmic behaviors. Participants were typically able to identify the kinds of biases that the research team had anticipated, and went further, pointing to additional potential biases in image search results. For example, while we gave the “wedding” search task due to its lack of representation of homosexual couples, many participants went beyond this to bring up other types of bias such as lack of representation of disabled people.

Potential Biases

In the next part of the interview, participants looked for new cases of bias and identified cases that were similar to the examples presented in the first part of the interview as well as entirely new observations of bias. For example, participants built on the “thug” search task to do similar searches such as “criminal” and “police” but also came up with searches like “Chinese food” that were unrelated to prior interview tasks.

In the diary study, participants submitted a total of 160 reports. 74 of these reports reflected observations that participants had made in the course of their everyday use of algorithmic systems, whereas the remaining 86 reports reflected discoveries participants made through active searching. The submissions described potential harm or bias spanning areas including gender, sexual orientation, race/ethnicity, socioeconomic class, age, body type, religion, politics, and combinations of these. 84 submissions included evidence or suggestion that the behaviors observed had to do with algorithms, and 126 included written evidence that the submitter thought the behavior was harmful. We will discuss many of the submitted cases in detail in the rest of the chapter; see Figure 4.1 for an example.

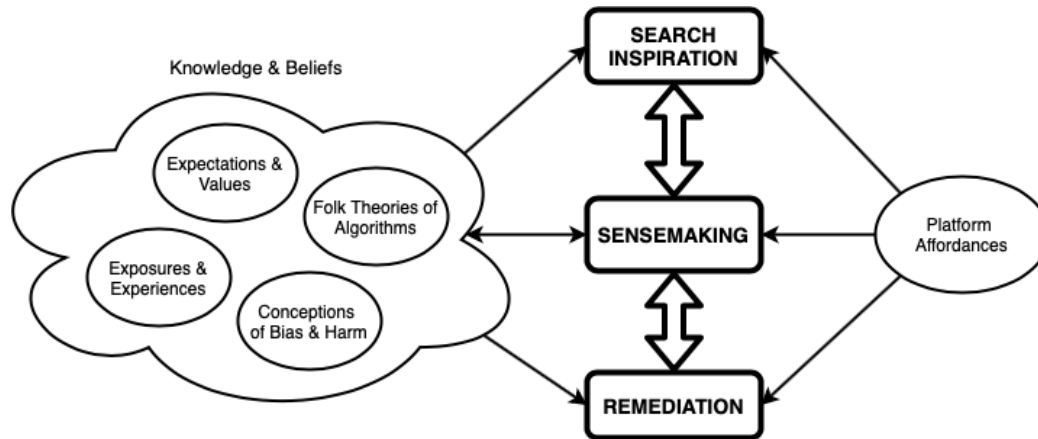


FIGURE 4.2: Process model depicting stages of and influences on users' bias search and sensemaking process. Bold boxes and text represent the three major stages of the process; other bubbles and text represent influences on and impacts of this process. Larger arrows represent movement between stages of the process, while the smaller arrows represent interactions between the process stages and influences on the process (i.e., platform affordances, user knowledge and beliefs).

4.4.2 Developing a Process Model of Bias Search & Sensemaking

Our analysis of data yielded a three-stage process model capturing the process by which people find and make sense of biased and harmful algorithmic behaviors as well as significant influences on different stages of this process. Figure 4.2 illustrates this model. In our open coding, three connected high-level stages of people's bias search and sensemaking process emerged: search inspiration, sensemaking, and remediation. While the process tends to occur in this order, it is not necessarily linear, and each stage can occur after any other stage. *Search inspiration* encompasses ways in which participants came up with ideas for where to look for potential biases in an algorithmic system. *Sensemaking* involves the ways that participants understood and evaluated algorithmic behaviors that might be harmfully biased. *Remediation* covers participant actions and desires to mitigate the ramifications of harmful algorithmic bias.

Connected to each part of this process, we uncovered common influences on the process across participants: knowledge and beliefs, and platform affordances. Participants used prior *knowledge and beliefs* about biased and harmful algorithmic behavior to make sense of potentially harmful algorithmic behaviors they encountered, and they gained new knowledge and beliefs along the way. Participant capabilities were made possible and guided by *platform affordances* such as aggregate views in interface presentations and autocomplete suggestions by search engines. In the next sections, we first present a high-level description of the influences, then delve into reporting of different stages of the process.

4.4.3 Influences on Bias Search & Sensemaking

In this section, we describe major influences on the bias search and sensemaking process that we observed in the three phases of our study: 1) users' knowledge and beliefs about bias and algorithmic systems and 2) the affordances for bias search and sensemaking that are present in a given algorithmic platform. In addition to influencing the bias search and sensemaking process, users' knowledge and beliefs are themselves influenced by the process via the sensemaking stage. Below, we describe

each of these influences and how they interact with users' bias search and sensemaking process.

Knowledge & Beliefs

Almost all participants (n=22) revealed that they held pre-existing knowledge and beliefs about biased and harmful algorithmic behavior that influenced the search and sensemaking process. In Figure 4.2, a bidirectional arrow is shown between sensemaking and users' knowledge and beliefs, reflecting that these knowledge and beliefs both influence and are modified via their sensemaking of specific algorithmic behaviors. Participants referenced specific kinds of relevant knowledge and beliefs: a) conception of bias and harm, b) exposure and experience, c) expectations and values, and d) folk theories of algorithms, each of which we describe below.

Conceptions of Bias & Harm The search and sensemaking process was heavily shaped by participants' conceptions of what constitutes "bias" as well as "harm". Participants defined algorithmic bias based on the intentions behind the design, the potential for imposition, and the priorities of algorithmic systems. Highlighting design intention, one participant defined bias as when algorithms *"don't actually do what they're intended to do"* (P7). Alternatively, focusing on the potential for imposition, another said, *"Bias means you are trying to influence the choice of the individual and try to impose your opinion"* (P18). Another participant defined bias by focusing on algorithmic prioritization, describing algorithmic bias as *"a preference towards certain individuals or certain languages or certain types of information, whether it's images or text, based on what is coded into the algorithm itself"* (P16). And some focused on the source of the bias, saying algorithmic bias arose via *"reflecting some human bias [...] injected through the data set or in the machine learning pipeline"* (P14) or via *"things that are calculated in the algorithm [...] through robotic systems that emulate humans"* (P22).

Participants' conceptions of harm also heavily shaped the search and sensemaking process. P3 mentioned a distinction between bias and harm, saying they *"see bias and negative impact and can't see harm."* In this vein, participants occasionally brought up cases that were algorithmic mistakes but not necessarily biased or harmful. For instance, although P14 would expect to see images of humans in the results of a search for "machine learning developer" (P14), the results mostly lacked visible faces and displayed abstract images and charts.

Some participants understood algorithmic harm based on the ramifications of algorithmic behaviors. A few participants saw harm as a primarily concrete and tangible effect. For example, participants described a situation where someone *"may not want to hire them"* (P16), which harms because *"it's a financial hurt"* (P10) and *"prevented someone from [...] employment, not being chosen, because of the way he dresses, the way he looks"* (P3). Others described how lack of representation *"further promotes the idea that you don't belong and perpetuates the sense of alienation"* (P2), how overrepresentation *"promotes a fairly narrow view"* (P12), and how misrepresentation *"reinforces a stereotype"* (P19). One participant teased out how harm might occur in some ways but not others: *"It's not going to cause any physical harm, but it could cause mental or emotional harm"* (P6). Another described a spectrum where harm is compared by severity: *"It doesn't really negatively impact someone to the extent that it's life threatening [...] harmful to a certain extent, but I wouldn't consider that superbly harmful"* (P21).

Other participants understood algorithmic harm based on the people affected by algorithmic behaviors. Some focused on those directly impacted. For instance, P1 described an image search for "thug" that showed predominantly Black men:

"It damages all of the Black community because if you're damaging Black men, then you're hurting Black families." Others focused on groups of people less directly connected to the algorithmic outputs: P12 mentioned broadly that *"kids are especially open to being influenced from stuff like this."* Many described algorithmic harm as being *"problematic for everyone"* (P19) because anyone's *"beliefs and views would and could be shaped by this"* (P8).

Exposures & Experiences Participants' prior exposures to and experiences of bias had large influence on how they went about searching for and making sense of biased or harmful algorithmic behaviors. In all the phases of the study, most participants (n=17) referred to a) their personal background and identity, b) their interactions with others, and c) external media sources when elaborating on the ways they had experienced and been exposed to biased and harmful behaviors. Below we describe these and their roles in the process of bias search and sensemaking.

Several participants (n=11) referenced their identity and the related experiences they had as members of particular demographic groups. For example, in the sense-making stage, participants drew upon their individual backgrounds to consider how image search might manifest harmful and biased behaviors. While looking at the image results for "group of people", P1 said, *"From being a plus-size person myself, I think about body positivity and body neutrality a lot."* This helped P1 decide that lack of body size representation in the results had harmful repercussions. Participants also referenced their personal backgrounds and identities while in the search inspiration and remediation stages. For example, when searching "weddings", P12 said, *"I'm bi and [...] would love to see more couples of women represented."* In this way P12 connected their identity to imagined ways to remedy problems showing in the search results. Participants also brought up their relationships and interactions with others in the think-aloud interviews and diary studies (n=11). For example, P1 reasoned about potential harm in search results after being tasked with finding an image of a thug in the interview: *"My partner's Black, and I just hate this whole stereotype that the face of crime is a young Black man."* Similarly, in the workshops, participants often drew upon prior exposure and experiences to relate to the user reports presented. For example, when discussing how the presentation of disability on search engines might generate harmful results in Workshop 3, P18 quickly contextualized the user-reported issue within a specific usage scenario that they had personally experienced (the diagnosis of an autistic child in their family): *"We needed to search on Google or other websites, depending upon the symptoms. And we are not getting the correct reflection from the search engine so ultimately if someone is searching... it may be misleading"*.

News, information about current and past events, and other mass media informed people's exposures and experiences of bias as well (n=12). Beyond movies and television, some reported in the diary study that they *"decided to read current news"* (P4) in order to find potential biases and drew in the interview from what is printed *"in fiction [...] or] in a newspaper or in a magazine"* (P14).

Expectations & Values Most participants (n=20) anticipated how an algorithm would behave in a particular situation, forming expectations about algorithmic behavior. In addition to predictions of how an algorithm will behave, participants also expressed normative expectations, or thoughts about how an algorithm *should* behave, which they linked to their general values. Both what participants expected an algorithm *would* do and what participants thought an algorithm *should* do in a given situation were important parts of the process, as participants evaluated ways that an

algorithm's observed behaviors diverged or aligned with those expectations. Participants began searching with expectations, then compared search results with their expectations to refine their conceptions of bias and harm (n=17). Sometimes participants expected and found bias (n=5). For instance, P12 conducted an image search for a "thug" and found overwhelmingly pictures of Black men: *"I've heard that the word 'thug' is generally racialized in addition to being kind of associated with delinquent or criminal activity, and so even just the history of that term kind of says— these results don't surprise me a ton."* On the other hand, sometimes participants expected bias but were surprised by search results (n=5). In the interview, P13 searched for "cooks" then "cooking" and said, *"I was expecting to see a bias towards females at home at least, which is obviously incorrect [...] yeah no, I think it's fine. It's not fitting any stereotypes."*

Folk Theories of Algorithms Most participants (n=18) mentioned folk theories about why algorithms behave in the ways that they do, which informed their search and sensemaking process. For example, one participant thought that *"Google has no control over what is seen"* (P10), placing responsibility to algorithmic behaviors elsewhere [167]. Another participant considered certain images in a set of image search results: *"Who is searching them to the point that they're so high up that that's what people have to see immediately?"* (P16). In this way the participant put forth a belief that other users' behaviors are impacting the ordering of what shows up while also expressing dissatisfaction with the result [480]. Participants displayed a wide range of algorithmic folk theories that interacted with their expectations [146], values, and conceptions of bias [288] to influence how they conducted their searches [165], how they made sense of search results [166], and what ideas for remediation they generated [146, 542]. This corroborates previous work that users develop folk theories to understand and guide their behaviors within algorithmic systems [145, 147, 165, 288].

Platform Affordances

All three stages of participants' bias search and sensemaking process — search inspiration, sensemaking, and remediation — appeared to be influenced by platform affordances. For example, the Google Images search interface presents an aggregate view of multiple image results in a grid as its default view, which makes distributional trends (and deviations from these trends) easier to recognize: when searching for weddings in a specific task, P1 only mentioned that *"these all seem to be very Western-looking weddings"* after seeing *"one is actually a different culture."* In other cases, Google Images' autocomplete suggestions inspired participants to search terms they might not have otherwise. One participant described a time when the suggestions caused them to be more precise in their search: *"As I was typing the word 'librarian', I could see a lot of the other choices that were popping up. And because I could see 'librarian memes', I didn't want memes, so that quickly caused me to modify my word"* (P19). Sometimes, Google's search query autocomplete suggestions inspired participants to propose related ideas to remedy issues they observed. For example, worried about being unknowingly influenced, P18 thought in *"that first initial searching, the user should be allowed to type completely through — no tooltip should be available."*

4.4.4 Stage 1: Search Inspiration

The first stage captures how people are inspired to search for potential biases and how people generate ideas for *where to look* for potential biases. This inspiration

can occur incidentally, during a user's regular interactions with an algorithmic system, or it may occur intentionally, when a person is actively trying to come up with inputs to an algorithmic system (e.g., search queries) that surface biased or harmful behaviors. Two main themes related to search inspiration emerged through our analysis: noticing patterns (or lack thereof) and drawing upon prior exposures to societal biases.

Noticing Patterns (Or Lack Thereof)

Sometimes participants noticed clear patterns or lack thereof in autocomplete suggestions, signalling the presence of potential biases and inspiring participants to look closer. For instance, P22 noticed when searching "nagging" that *"'wife' immediately comes up"* many times in the autocomplete suggestions. As another example, P23 noted during the diary study that Facebook stories have automatically suggested reactions saying "beautiful" that *"only popped up on stories where there was a woman or feminine figure featured or a pet."* Similarly, when participants noticed discrepancies between their expectations versus observed algorithmic patterns, they theorized about potential biases, and tried to look deeper. For example, P21 noticed and submitted a potential bias in the diary study *"because it was an advertisement for entertainment news and I do not follow any celebrities on my Twitter account"* and, similarly, P23 said, *"It struck me because it is so out of character for the folks I follow."*

Drawing Upon Prior Exposures To Societal Biases

Participants were often inspired to search for particular kinds of algorithmic biases based on their own prior experiences or exposures to societal biases. For example, when brainstorming during the open-ended interview task, P12 said, *"I study computer science, so I'm sure if you google computer scientists you would get a lot of men."* One participant explained that searching "image of a maid" was inspired by the racial bias depicted in the movie *The Help*: *"I thought I was gonna see some ladies from The Help [...] that was my first thought when I chose to search this particular image"* (P3). This participant drew from the movie's representation of Black women as maids in the movie. And P1 referred to a discussion with their partner about under-representation of women in the video game community as an inspiration: *"My partner really likes playing video games and we talked a lot about how the video game community skews male [...] so I'm going to try to look up gamer because that might pop up some things."*

4.4.5 Stage 2: Sensemaking

In the course of testing various inputs to algorithmic systems, participants inspected and made sense of algorithmic outputs to evaluate whether they might encode harmful biases. This sensemaking process generated new inspiration for searching, new beliefs about algorithmic systems, and new ideas for ways to remedy problematic algorithmic behaviors that participants observed. Below, we describe the related themes that emerged in the sensemaking stage.

"The Bias Is in the Eye of the Beholder": Interpreting algorithmic biases through the lens of their own biases

Participants made sense of potential algorithmic biases through the lens of their own stereotypes and prescriptive desires. The fact that *"everybody has their own definition*

about [...] bias" (P4) led participants to subjective interpretations of whether what appears in search results is biased/harmful or not. In one case, many participants mentioned that the concept of "beauty" is subjective (n=6), so it is unclear how a search should represent that. For example, P1 said, *"I definitely do have more of that kind of Western society bias [...] I hadn't really thought about that kind of cultural aspect of it as much."* Similarly, when P10 was evaluating the search results of "beautiful woman" for potential biases, they did not consider the presentation of beauty via makeup a harmful bias because *"I feel like a woman should wear makeup. [...] I guess that's the way my mother feels, so it kind of drips down to me."*

Different conceptions of bias among participants sometimes prompted opposing views of whether an algorithmic output encoded harmful biases or not. For example, after searching "beautiful woman", P8 said, *"I was pretty sure [...] everybody would probably be represented here"* and concluded that *"everybody like I said is represented here as far as race goes."* However, many other participants commented on the lack of racial representation when they performed the same search task: *"They do need more diversity [...] I don't see anybody with different skin tones"* (P7).

The Interaction between Bias Evaluation & Perceptions of Reality

Participants compared algorithmic outputs with their perceptions of reality to evaluate what is and is not biased (n=13). For instance, P13 considered the results from a specific task to find an image of a wedding and said, *"As an Indian, I know that it's not accurate for India"*, comparing their experience of reality to help them examine potential cultural bias in the search result. Some participants believed that results that reflected reality were necessarily less biased: *"the demographics fit the population which is what I think it should come out to be when you do a search."* (P5). In practice, this led to the expectation that search results should reflect reality: For example, P5 thought a search for "professor" during the interview might not be biased despite showing little gender and racial diversity because *"maybe most professors are white males, in which case I think it would be fair that's what the search shows"*, or similarly, *"If all firefighters were white men, then that's I think what should come up"* (P5).

Some had more nuanced insights when comparing with reality, saying that, though a search might technically be accurate to real-world data, *"the data might not be reflecting the reality"* (P14). P14 provided an example: *"If you look at the surface there are indeed more African Americans being arrested or reporting violent [...] but then this is institutional bias, right?"* In this way the participant elaborated that data is not a neutral reflection of reality and can contain biases as well. Similarly, some thought the results were *"maybe not necessarily inaccurate towards how [a search term]'s used. It's just, the way it's used itself is kind of biased"* (P12). In line with findings from [256], a few participants noted that some biases might reflect reality in informative ways, and preferred to preserve these biases: *"It's reflecting human condition and human bias, which is actually good because if you make it extremely balanced, it's almost unreal"* (P14) and *"it would seem almost wrong if it was all reversed"* (P5). However, it has been pointed out that real-world search results can conflict with outcomes desirable to society [294].

While participants' perceptions of reality impacted how they interpreted and evaluated potential biases of algorithmic outputs, they were also shaped by algorithmic outputs. A few participants (n=7) who believed representation in search results should reflect reality concluded that the search results were likely accurate *regardless of any biases present*. In line with prior literature on perceptions of bias in image search (e.g., [294]), these participants then used search results as a basis for

making inferences about reality. For instance, P10 said about a search, *“I see a lot of women, not a lot of men”* and concluded, *“I guess it’s a predominantly women’s field.”*

Reflecting on Uncertainty in Bias Evaluation

While participants tried to interpret the algorithmic outputs in various cases, several (n=10) described difficulties in knowing whether biases were actually present. Repeatedly, participants mentioned that they *“really didn’t have any solid evidence, it was just a feeling”* (P3). Participants pointed out that the information available to them was insufficient to prove some of what they noticed. For example, when looking at image search results for *“beautiful woman”*, P19 said, *“I also don’t see anyone who would jump out to me as a transgender woman, and so I do believe that this leaves them out. And I mean, sometimes it’s hard to tell.”* In this way, P19 detected a potential lack of representation while simultaneously acknowledging that the way someone presents does not necessarily have bearing on that person’s gender.

Reflecting on their uncertainty, participants discussed the information they would need to have a more certain evaluation of an algorithmic outcome. For example, in the course of their collective investigations in the workshops, participants often reflected on missing contextual information needed to properly evaluate a given user report. In some cases, participants recognized that they lacked crucial information needed to determine whether a reported issue represented biased and/or harmful algorithmic behavior. For example, in Workshop 1, when discussing the pierced earrings case (refer to Table 4.1), P5 expressed uncertainty that the reported behavior actually resulted from algorithmic decision-making: *“I don’t know [...] I was under the impression it was an ad from a company, you know it wasn’t a Google search, so I just thought, maybe that company sells that specific style of earrings (for women) [...] men wouldn’t wear them.”*

Workshop participants sometimes asked specific relevant questions about information they lacked for a reported case. For example, in the Workshop 4 group report, when trying to assess whether a user report provided evidence for a possible gender bias in the YouTube recommendation algorithm, participants reflected on the missing information needed for an effective investigation: *“Does the user have personalized ads on? Their actual gender and age? Incognito mode or not? Prior search history do they often search for this kind of content?”*.

Advocacy & Solidarity

As participants reasoned about algorithmic harms, they sometimes advocated for marginalized demographic groups. For example, during Workshop 1, as participants discussed the limited diversity in search engines’ representations of romantic couples, P10, middle-aged at 46, argued that elderly couples would be harmed: *“When my grandma said she had a boyfriend, [my little cousin] was like, ‘Grandmas don’t have boyfriends.’ [...] It’s just a stereotype and it’s ingrained in people’s memory.”* While advocating for marginalized groups, participants sometimes discussed the further consequences of such algorithmic biases: P16 in Workshop 2 said, *“By whitewashing these disabilities, whether they’re visible or not, it also leads to some of the issues in the medical field where doctors don’t trust women and also minorities and people of color when they talk about their pain threshold.”*

Participants’ advocacy of marginalized groups sometimes turned into collective solidarity during workshops when participants worked together. For example, we

observed that by drawing upon prior experiences and invoking parallel cases, participants sometimes attempted to form a shared identity with others in their workshop groups. A particularly interesting example emerged during Workshop 1. Although the group consisted of participants with different self-reported demographics, including two women and one non-binary participant, participants seemed to form a shared femininity in their discussion of the pierced earrings case. For instance, in the Workshop 1 final report, the group elaborated on why this was harmful: *"Society dictates how we act as women and how we should dress and wear jewelry. It is kind of expected by others that we wear these things in public."*

Nuances in Meaning of the Selected Algorithmic Input

Participants also considered the various ways their inputs to algorithmic systems could be interpreted when considering potential bias (n=7). Some noticed linguistic connotations because search results depicted a singular version out of many possibilities: P2 image searched "elegant" in the diary study and said, *"It almost exclusively depicts very thin, white women in expensive evening wear [...] 'elegant' can just as easily apply to a math concept or sculpture."* On the other hand, some participants noticed linguistic elements because search results depicted more possibilities than what was imagined while searching: P11 image searched "model" in the diary study and said, *"interestingly, one of the results is a biomedical model of the skin [...] I didn't stop to think about other semantic interpretations of the word 'model'."* This could lead to participant analysis of word meanings, as participants considered how definitional and interpretational differences between terms might affect a search. For example, when image searching terms like "fashion", "cute dress", and "sexy" in the diary study, P17 found mainly images of white women and said, *"These words even though they are not gendered seem to have heavy gendered interpretation."* Participants wondered about how some words can have more neutral connotations, how other words *"are imbued with a harmful, [racist] connotation, and how this arises in images online"* (P11). Some mused about how communities reclaim certain words used negatively to describe them and what that means for image search representations. For example, P4 described how "thug" could be highly associated with Black men in search image results partially due to some Black men rappers choosing to self-describe as "thugs". P19 mentioned that "thug" could have been reclaimed in the same way as "queer", *"a word that has been reclaimed and chosen by the LGBTQ community"*.

Sensemaking through Comparison

Many participants used comparison on the platform as a method for evaluating and making sense of search results (n=14). Sometimes this meant doing more searches as a sort of test. For example, P16 image searched "good neighborhood" and then, in order to make sense of the search results, searched "bad neighborhood", comparing the visual depictions of the two. Similarly, participants followed up image searches for "nagging" with searches for "men nagging other men", "nagging man", and "man nagging" in (unsuccessful) attempts to force depictions of men who are nagging into the search results (P10, P14).

Other times, this meant comparing the amount and type of representation. In Google Images, this could take place across multiple searches, as participants used the amount of representation in one search result to situate themselves in another, or in a single search, as participants compared resultant images to each other, something Google Images' search interface makes straightforward. Perplexingly, noticing

one or two images of a type of person sometimes led participants to conclude that the ratio of representation seemed bad within the set of results and thus the search was biased, but other times led participants to conclude that various types of people were represented and thus the search was not biased. Past work has discussed the role of comparison in sensemaking and other cognitive processes [210].

Placing Responsibility for Algorithmic Outputs

Participants considered what controlled platforms and caused bias to appear. Some participants believed that Google intentionally produced the representation — or lack thereof — in image search results (n=4). When more diversity appeared in a search result, participants expressed that *“they’re trying to achieve better representation”* (P12). Participants left *“they”* unspecified, a pronoun with no antecedent, some person or people who have the power to unleash their intent in Google Images search results. When less diversity appeared, participants similarly claimed that *“there’s a pretty narrow definition being pushed here”* (P12).

Others theorized that the algorithm itself was neutral and that results were a reflection of the data (n=6). Explaining the lack of diversity in a search result, P10 said, *“They might not have the picture available”*, moving responsibility from the algorithm to the data. Others posited that Google was *“not hand picking this stuff”* (P12), blaming their own search behavior for their inability to find better representation in a search: *“It’s probably because I’m not searching [...] this is just a brief search. If I took the time, I’d probably find some”* (P4). This suggested that search result bias is symptomatic of the searcher’s bias in selecting certain search terms in which they could *“demonstrate something without thinking”* and inadvertently neglect *“other things that [they] could choose instead”* (P11). Others blamed results not on individual users but on users en masse, claiming that results were *“based on what people have been typing into the search bar”* (P21).

Many participants thought the algorithm targeted individual users (n=8). Participants described this as personalization that provides *“the response that’s best suited to the question you’re asking”* (P16). Some participants thought this happened via *“targeting advertising based on the metrics that I provide”* (P23), leading to a situation where, for example, *“they think I’m a 40-year-old man because of my athletic preferences or something because I clicked on some shoe thing”* (P11). Another theory suggested that the search results changed based on time relevance: *“I think it’d be interesting to see how the algorithm changes with current events because it’s definitely something that would change on a day-to-day basis”* (P11). Many hypothesized that results were *“probably based on [their] location”* (P1) since *“Google definitely knows where I am”* (P13), which led participants to *“wonder how this changes when you’re looking at [results] from different countries”* (P16).

We similarly observed in the workshops that, in the process of collectively investigating whether an individual user report represented a case of harmful algorithmic behavior, participants questioned how the behavior originated. However, they often had a hard time untangling different possible explanations for an observed behavior based on the information provided in the user report. For example, in Workshop 2, when discussing the disability case, participants started a discussion to figure out the sources of bias and eventually concluded that it could be a mixture of multiple factors: *“It could be algorithm or human... or sometimes it’s both? Even if there was a developer, putting in the source code in the beginning, if input [is biased...] the computer will be running on that additional information”* (P16). Responding, P21 said: *“It’s a couple of different factors. Like how much people are searching, that could be one of the factors”* (P21).

4.4.6 Stage 3: Remediation

The third stage describes the actions that people take to avoid or counter harmful biases. We observed that, while reflecting on potential changes that could be made at a platform or societal level to counter harmful biases, people often reinterpreted the algorithmic behaviors they had observed, questioned whether these truly represented biased algorithmic behaviors, and reflected on relevant information that they lacked, leading them back to sensemaking. Through sensemaking, some remediation actions also informed search inspiration. For example, in the context of image search, we observed that some people avoid searching for a particular term or alter their search terms in efforts to avoid seeing potentially harmful biased search results.

Participants valued remediation of potentially biased and harmful algorithmic behaviors (n=22). Through their actions and desires, they expressed ways to try to fix potential issues they encountered. When they were able, participants took action; other times, participants described methods for the platform or society more broadly to take action. Five types of remediation emerged — representation, resistance, awareness, pressure, and communication—and are described below.

Improving Representation in Content, Diversity, & Ordering

Improving representation included desires for increased diversity and for increased accuracy in representation (n=17) — that is, wanting *“to get accurate results and [wanting] to get diversified results and [not wanting] to use stereotypes or certain biases against certain types of people”* (P22) — goals that are sometimes, but not always, aligned. This tension highlights that bias removal is not always a reasonable option, as people have different values, expectations, and conceptions of bias. Beyond broadly thinking that results *“should represent a more diverse group”* (P17) or should *“be more statistically correct”* (P13), participants described concrete methods of curation by which they thought these representation aims could be achieved.

One described method involved bringing some of the images *“up to the top”* because *“that might be more balanced than what pops all the way to the top”* (P19). Re-ordering images to *“add in more [diversity] in the first two rows”* (P21) would ensure that representation appears *“in the beginning, where people most are paying attention”* instead of further down the page, *“where people aren’t going to be scrolling”* (P9). A related method put forth that *“it’d be good to have one [image] after the other kind of like alternating with different faces, different ages, different ethnicities and races, just to make it clear that everyone’s represented here”* (P16). To achieve this curation, some pointed to a platform-level need to *“be more careful about balanced data set”* (P14) and even consider *“[doing] it in a way where [platforms] have their own images”* (P10) instead of pulling from other sources for data. With or without curated results, P18 emphasized that users can always *“go more deep to the search or go to the next page”* to get more representation.

Individual User Resistance

Participants took a variety of actions to counteract bias themselves (n=12). Some created representation themselves out of inadequate results by *“finding images [...] that don’t perpetuate those stereotypes quite as much”* (P12). When potentially biased results appeared after using a search term, others opted *“to not search for that [term] and try to search for something else”* (P7), to *“try different word searches in the search box”* (P4), and to *“be more specific in the way I search”* (P16). For example, after noticing a

dearth of cultural representation in a “wedding” image search, P13 said, “I’d type in ‘Indian wedding’ and [that] will definitely be what I would be expecting for sure.”

For some participants, being reactive was not enough: they sought preemptive remediation. When tasked with finding an image of a wedding, P11 searched “lesbian wedding”, explaining, “I’m in a lesbian partnership right now and wanting to get married and envisioning a wedding [...] and I’m so sick of [seeing] these straight weddings.” Similarly, when tasked with finding an image of a beautiful woman, P19 searched “women”, explaining, “I was concerned that if I typed in ‘beautiful women’ I might not like it. I might get a whole lot more white people, and I might get more swimsuit models and stuff like that.” In this way, participants showcased their awareness of biased norms, highlighting that at some level they knew they needed to alter their search terms to try to guard against this and get the desired results.

Other participants took even stronger measures. Some attempted to prevent their ability to see potentially harmful content by trying “to block out the stuff that offended me” (P10) and “to avoid that thing, block that particular person or group or website” (P18). And one participant said bias could make them “reconsider the kind of mediums that I’m using, the kind of information I’m taking in” (P16). Such attempts to resist algorithmic bias align with the concept of everyday algorithmic resistance [542] in which users try to exercise their agency over algorithmic structures. Past work also showed users resisting algorithms to define their own identities in relation to the algorithm [288].

Fostering Awareness of Bias

Participants sought to bring bias to the surface of people’s minds through raising various kinds of educational awareness (n=16). Sometimes this awareness raising occurred at a societal level. For example, P14 said that a search result was “so bad that it’s actually good for society because it just reminds people how this kind of stuff has happened before [...] we should really dive into this phenomenon and look into the root of the discrimination and try to change it.” In this way, P14 suggested that demonstrations of bias could spur further anti-bias work.

Some thought that users should be “trained how to use the technology [...] what are the do’s and don’ts, what are the ethics” (P18) and taught “how to do more robust image searches and more robust searches” (P19). Some particularly suggested public education. For example, in Workshop 2, when discussing how certain forms of disability (e.g., depression) are rendered invisible in image search results, participants advocated for public education around the nature and scope of disability as a path toward addressing this: “They [the public] need to be educated about what a disability is” (P19). Similarly, participants in Workshop 3 wrote in their group report that “A lot of regulations around [accessibility]... people are just not aware of it. There are a lot of organizations to go for. It’s well researched and documented but not getting enough attention from the population. [We need to] survey the public about their awareness.”

Additionally, participants thought that the tasks they were asked to do as part of the think-aloud interviews and diary study could be useful in educational contexts. P11 described being able to “use [these] as teaching tools a lot. I teach a lot about harmful stereotypes [...] like we do semantic prototypes for one of our topics,” and P12 thought that the tasks were “super valuable in terms of portraying [algorithmic bias] pretty clearly” and could be useful to “teach about algorithmic bias in some kind of interactive thing.” Participants felt the tasks were educational for themselves as well: “It was very important, revealing just the disparaging of minorities on a daily basis, in certain places” (P8).

Some participants said that they would raise awareness on a more personal level. Telling others who think and feel similarly presented as a method to both raise awareness and share feelings. For example, P1 said, *"I feel like my friends and I will commiserate on it, since it's something that several of us experience."* Similarly, P22 would *"share with my friends to kind of laugh at why is this the first thing that shows up?"* Telling others who might not think and feel similarly also came up. One participant did this *"to see what their opinions are"* (P9). Another *"[raised] these points to my peers who may not be as aware to start the conversation"* (P16). And others described doing this to teach others who *"could probably benefit from hearing that discussion"* (P1). For example, P12 said, *"I've got a little sister. I would probably be like, 'Yo, these search results are kind of bad. Here's why.'"*

Exerting Pressure on Companies

During our workshops, participants expressed desires for companies to take responsibility and fix issues on their platforms. Collectively, they described how they might pressure companies to take action using the aforementioned awareness raising. For example, in Workshop 4, P13 wanted to give the report to *"a news agency, because I think if it went straight to the developers they might ignore it if it's something that makes them money."* In this way, raising awareness connects to influencing a company via publishing issues to a wide audience. As P9 also in Workshop 4 said, *"You want the news to get out to the whole public and reach a wide variety of viewers."* Similarly, P16 in Workshop 2 would *"recommend reaching out to civil society and maybe nonprofits that [...] would have a stake in this and sometimes they can put more pressure on larger companies like Google."* As a concrete way to exert pressure, P19 in Workshop 2 brought up money, saying they *"could also pick a couple advertisers who would not want to be associated with those biases and see if we could get their attention on it, [see] if we couldn't get Google to act."*

Communication between User & Platform

Participants thought platforms could communicate with and understand users in ways that better address harmful algorithmic behaviors (n=15). A frequently mentioned method to accomplish this involved changing the communicative burden [331] to rely less on user behaviors. Some participants wanted the platform to understand users via idealized personalization that could *"magically come up with the thing that I want in my head"* (P11) so that users *"get the response that's best suited to the question you're asking"* (P16). However, participants knew they could be *"so biased by everything that it pops up"* (P19); because of this, one remedy proposed that *"the search engine should allow the user to type the complete criteria or word or sentence [they] want to search"* (P18).

Others wanted the platform to provide additional information about what was happening. Some thought explanations of algorithmic behavior and how to interact with it could help: P21 thought *"a note [...] explaining why certain searches give you certain results could be useful,"* and P18 similarly said that platforms *"should provide a module or instructions, do's and don'ts on the top."* Though this action would be done by platforms, it would presumably result in changed user behavior that leads to less biased results. Others wanted platforms to provide more context about some searches, such as by offering users *"all of the different meanings [of a term] up in the first screen"* (P19) or through *"a feature that allowed you to get a description of the term before*

you see images" (P16). Another mentioned way to include context was through *"some kind of disclaimer saying, 'This word has a racialized history'"* (P12) in relevant places.

Participants also would appreciate increased ability to give feedback to platforms. Wanting to report a concern, P3 wondered, *"Who do I write to, to make this complaint?"* In that vein, P18 said, *"If [the] option is available, then I'll submit [issues] as spam or harmful post."* Others described a slightly different interaction style where *"you should be able to flag [issues] for someone to either remove or add context"* (P16) and *"mark them as inappropriate"* (P13). Participants emphasized the need for ease, as they are *"not going to go out of my way to look for people to tell"* (P8), and for action to be taken in response, as P11 described situations—*"when I tend to report harmful things on Facebook or Instagram [...] it still doesn't go anywhere"*—that led to frustration with ineffective reporting processes.

4.5 Limitations

It is important to highlight some specific limitations of our methods. In this study, we focus on image search as our problem domain for the interview study, as it provides a familiar, straightforward starting point for our participants to search for problematic machine behaviors. It offers an interface that supports easy comparison across different images, through which biases in result ranking and distribution can be more easily identified. However, biases are also present in other types of algorithmic systems and digital platforms, where such patterns may not be salient. In addition, this kind of comparison-based pattern detection is not always relevant to identifying harmful machine behavior, which can manifest as a single problematic instance (e.g., an image-caption pair) rather than a statistical pattern [231]. Also, when considering harmful bias, a large number of participants brought up biases that were caused by misrepresentation and stereotypes. This focus might have been influenced, at least in part, by the examples we chose in the specific interview tasks and the use of image search as our initial problem domain in this study.

4.6 Discussion

Past literature has shown that everyday users can be powerful in detecting harmful and biased machine behaviors in their daily interactions with algorithmic systems [474]. However, little is known regarding how such power can be supported and guided. In this chapter, we have begun to shed light on these questions by investigating the strategies everyday users use to surface potentially harmful algorithmic behavior, both individually and collectively. Below, we discuss design implications and future research directions toward supporting user-driven algorithm auditing.

4.6.1 Supporting User-Driven Algorithm Auditing

In this chapter, we investigated how people surface harmful algorithmic behaviors with the aim of informing the design of new tools and methods to support user-driven algorithm auditing. We presented a process model capturing influences and stages of users' bias search and sensemaking process. Our investigation primarily focused on scenarios in which participants *actively* search for potential algorithmic biases; future investigations could explore how we might better support people in following up on incidental encounters with algorithmic biases. Here we discuss how

our process model can inform the design of new interventions to facilitate future user-driven audits.

Utilizing Prior Experiences & Exposures in Uncovering Harm

The results of our studies suggest that people's exposures and experiences have a large influence on what types of bias they are able to identify. In addition, beyond their personal identities and backgrounds, participants drew upon second-hand knowledge of the experiences of those close to them, or knowledge from media that they had consumed. Assigning people to particular auditing roles based on their specific exposure and experience could boost future auditing efforts. For example, in order to surface potential racial biases in an algorithmic system, it may be most useful to engage users with exposure to and experience with related racial bias. Further, this indicates that having auditors with a diverse set of exposure and experience is crucial to comprehensively unearth harmful algorithmic behaviors.

Participants also showcased their own cognitive biases as they made sense of potential biases. Sometimes participants entered a term and said that they searched it because they expected a certain bias to appear. Then, despite having little to no evidence to support the expectation, they concluded that their expectation of bias was correct. It may be that people held preconceptions so strongly that they fell prey to confirmation bias. Future supports for user-driven algorithm auditing should anticipate such cognitive biases, and should be explicitly designed to mitigate their impacts.

Facilitating Bias Detection & Reporting via Platform Affordances

In our studies participants mentioned desires and actions to raise awareness of problematic algorithmic behaviors. Participants described telling other people that they know, sharing information with impacted groups, and getting the word out via mass media, and they also expressed that they would like to communicate issues with platforms and companies. We can leverage our knowledge that platform affordances influence the process by providing straightforward ways to report bias, either to the platform or beyond, through the design and mechanisms of the platform itself.

We also know that participants frequently used comparison as a method by which to notice, understand, and evaluate potential bias. For instance, Google Images' search interface allows users to easily view groups of *multiple* images in ranked search results, which is useful for examining patterns and finding biases that only emerge through comparison. Participants also used comparison to look for more evidence, testing for the presence of bias through levelled searches that they would then compare to each other. This proved helpful to participants' consideration of bias. Employing platform affordances to facilitate this comparison could lead to richer and more fruitful user-driven algorithm audits.

Another type of platform affordance comes in the form of quantification. Although we see potential for user-driven auditing in identifying qualitative harms, participants sometimes attended to quantitative information present in the platform in forming their judgments. As one example, participants often looked for proportions represented in image search results, which could be productive in surfacing certain issues where participants noticed a dearth of certain categories. But we also observed several cases where we believe it was distracting to look at proportions, such as when participants identified the presence of a single instance of a category

as proof that there was enough representation. So quantitative reasoning for qualitative harms can be helpful or counterproductive, depending on the method.

Toward Building Algorithmic Counterpublics

As previous literature [474] suggests, when a group of users work together to collaboratively make sense of problematic algorithmic behaviors, they often form what Nancy Fraser termed as “counterpublics” [186] – where members of often marginalized social groups collectively participate in their own form of sensemaking, opinion formation, and consensus building. For example, via the platform’s internal discussion forum, Yelp users were able to form algorithmic counterpublics against biased rating algorithms [169].

We observed similar counterpublics forming in our four workshops. Indeed, in the workshop sessions, we saw how participants from various backgrounds attempted to establish common understanding around the issues being reported via actions such as untangling the sources, reflecting on the context, and proposing possible remediation solutions. How can we better support such collective auditing behaviors to build more successful algorithmic counterpublics?

We observed that the lack of contextual information remains as one of the major challenges for our workshop participants, as they often had a hard time untangling different possible explanations for an observed behavior based on the information provided in the user report. One way designers might help auditors overcome this challenge is to allow them to ask follow-up questions of people who report their observations. For example, since none of these workshops had access to the demographic information of the user report submitters, participants were constantly questioning whether the problematic machine behavior could be a type of “demographic targeting”. Future designers should consider strongly encouraging auditors to provide necessary contextual information (e.g., demographic information) to better facilitate such collective sensemaking. Another way for designers to support such types of collective sensemaking is to offer cues to remind the auditors to reflect on the missing information. For instance, in workshop four, participants did not necessarily reflect on missing information until filling out the final report, after they had already discussed a given case at length. In this case, some of the questions in our workshop’s final report served as cues for our participants.

We also observed interesting dynamics of the demographic makeup of our workshop participants. For example, if in a workshop, all participants shared similar backgrounds (e.g., shared experience or similar technical expertise), they might be missing certain information in bias validation due to the constraints of team diversity. However, sometimes similar backgrounds and shared identities might also allow people to bring in new ideas without them seeming threatened. Future designers should consider how to balance such dynamics in supporting more effective team formation.

There is great opportunity in encouraging users to come together within their auditing activities. Though users can and do individually audit algorithms, oftentimes it is only together that can they conduct tests as different, authentic users to gather robust evidence. Only together can they discuss and provide general consensus or disagreement, or raise awareness at the scale required to put pressure on companies and hold them accountable.

4.6.2 Limits & Challenges of User-Driven Algorithm Auditing

Though user-driven algorithm auditing has many strengths, it comes with its own challenges and limitations. Everyday algorithm audits have had varying levels of success [474]. One relevant factor may be the difficulty in creating productive structure and organization in a user-driven algorithm audit. Establishing structure and organization is especially difficult amongst large groups of users, though collective action is exactly what many user-driven audits need to interrogate, hypothesize, test, and raise awareness effectively. Our process model offers insights for how users work, together and independently, that can be leveraged to carefully create and support this structure and organization at scale.

User-driven audits thus far have occurred in situations where users notice potential issues within algorithmic systems, whether by stumbling upon or by actively searching for problematic algorithmic behaviors. This limits the relevance of user-driven auditing to algorithms that are directly visible to users to interact with, neglecting many algorithmic systems that are offline and invisible to users. However, tools and platforms that facilitate user-driven auditing can be designed to provide and encourage interaction with algorithms that users do not normally interface with. A similar but distinct challenge involves situations in which those targeted and potentially harmed by algorithms are not the direct users who interact with these systems day-to-day. For example, in many high-stakes algorithm-assisted decision-making contexts like child welfare decision-making and recidivism prediction, public sector workers are the direct users of such systems, but the citizens and usually marginalized people are the ones who are impacted by these algorithmic systems' outcomes. In such settings, it may be beneficial to bring in diverse stakeholders with relevant knowledge—especially those who are impacted by algorithmic outcomes.

Not all algorithmic behaviors are equally visible, and sometimes people are harmed by algorithmic behaviors without their knowledge. In this chapter we have largely focused on harmful algorithmic behaviors that people have opportunities to encounter in their day-to-day lives and that (at least some) people can plausibly be expected to recognize as harmful. Not all algorithmic harms fit this description. In some cases, this may be because those most affected by a given algorithmic system are not the users who directly interact with the system, as discussed above. It may also be due to the opaque nature of some algorithmic systems. For example, gig workers often are unable to tell how the algorithm managing them works. But apps to compare wages have helped gig workers to collectively reveal inequities [352]. The creation of new platforms like these apps, with affordances that support collective investigation and sensemaking, can provide avenues for the recognition of harmful algorithmic behaviors that may otherwise remain invisible.

Determining what is noise and what is legitimate, relevant information is another challenge. Often the users are not technical experts and have no straightforward way beyond their own input/output based testing to determine the correctness of their hypotheses about how the algorithms work. Identifying signal versus noise is recognized as a challenging problem for user- and crowd-driven approaches in general. Some bug bounty programs use experts to ensure they have more signal than noise [218]. But weeding out noise requires care, as the process can involve a level of subjectivity. Apparent “noise” may actually be an important signal, indicating where the ground truth is contested (e.g., representing substantive disagreements across different social groups) [351, 451].

Of note, *supporting* user-driven auditing inevitably involves priming users' perceptions and auditing behaviors to some degree. Just as we primed participants with

specific tasks and examples in this study, users on a platform would also necessarily have to be primed to some extent in order to support the process of user-driven auditing. For example, the onboarding or training provided to users who are new to a user-driven auditing platform may influence their subsequent bias search and sense-making behaviors, no matter how diligent platform designers are in reducing such priming effects. Similarly, if a user-driven auditing platform recommends specific auditing tasks to users, there are risks that the selection and presentation of tasks will inadvertently serve to prioritize the interests of some social groups over others, or that the structure of a given task will facilitate identifying certain kinds of biases over others.

4.6.3 Algorithmic Harm Remediation: What's next?

A critical question that any user-driven auditing platform must be designed to address is this: in cases where users identify harmful algorithmic biases, what might it actually look like to “fix” the issue? Our participants showcased varied definitions of bias through their differing interpretations of algorithmic outputs. Some participants envisioned accuracy and matching with reality, while others envisioned diverse representation regardless of accuracy. For accuracy, the process of choosing ground truth is fraught, as many sources are littered with biases and other issues themselves. For diverse representation, what would ideal representation look like? Interestingly, participants found representation important in both positive (e.g. “beautiful woman”) and negative (e.g. “nagging”) contexts. This begs an ethical question about when the negative context is harming those who are represented and thus increasing representation would consequently harm more types of people: Is representation the answer? And if it is, who or what should be represented in negative contexts like this? These tensions would need to be dealt with in any future remediation.

In many cases, it may not be possible to simultaneously achieve all remediative ideals, especially when there is no alignment as to what bias is in the first place. This means that designers of future user-driven auditing platforms cannot fully escape making design decisions that will impact *whose values will be prioritized and how*. The design of equitable processes for recommending or deciding upon remediation strategies is a critical open problem in this space.

When fixing bias is neither desirable nor possible, promoting awareness of algorithmic biases and other harmful behaviors can be especially valuable. User awareness helps users adapt their behaviors in and around algorithmic systems, in turn fostering user-driven auditing and supporting remediation of harmful issues by other means. Participants described a wide range of ways that awareness of bias could be promoted, and most of the other remediative methods they described require user awareness in order to commence.

Beyond discussions of how algorithmic systems should be “fixed” are questions about how many outstanding issues we as a society are willing to accept and mark as ‘need to be fixed’ before taking more severe recourse. Other researchers have highlighted the ways that algorithmic harm derives from the structural power given them [22, 364]. Though a common reaction to harmful algorithms is to address specific issues, there have been times when the response has instead removed the algorithm’s power entirely (e.g., [136, 307, 549]); indeed refusal instead of repair has been highlighted as an often appropriate way to respond to bias [53, 568]. At what point in a harmful algorithm’s existence do we mitigate the harm by stopping the algorithm altogether rather than just attempting to mitigate specific issues?

4.7 Conclusion

In this chapter, I demonstrated everyday users' strong abilities to detect and reason about potentially harmful algorithmic behaviors through a three-stage process model (i.e., search inspiration, sensemaking, and remediation) as well as significant influences on the process (i.e., knowledge and beliefs, and platform affordances). These findings reveal many opportunities in designing user-driven algorithm audits such as utilizing users' prior exposure to and experiences of harms and biases as well as leveraging platforms' affordances for comparing and reporting potential cases of algorithmic harms. The collective effort of users in the workshops to understand, investigate, and consider remediations for potentially harmful algorithmic behaviors also showcase opportunities for building counterpublics and collective audit platforms. However, balancing the dynamics of such platforms including the demographic makeup and community guidance and feedback needs further research, along with a careful design process.

Chapter 5

Understanding Practices, Challenges, & Opportunities for User-Engaged Algorithm Auditing in Industry Practice

This chapter is adapted from my paper [141]:

Wesley Hanwen Deng, Bill Boyuan Guo, Alicia DeVrio, Hong Shen, Motahare Eslami, Kenneth Holstein. “Understanding Practices, Challenges, and Opportunities for User-Driven Algorithm Auditing in Industry Practice.” In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. (2023).

Recent years have seen growing interest among both researchers and practitioners in *user-engaged* approaches to algorithm auditing, which directly engage users in detecting problematic behaviors in algorithmic systems. However, we know little about industry practitioners’ current practices and challenges around user-engaged auditing, nor what opportunities exist for them to better leverage such approaches in practice. To investigate, I conducted a series of interviews and iterative co-design activities with practitioners who employ user-engaged auditing approaches in their work. The findings reveal several challenges practitioners face in appropriately recruiting and incentivizing user auditors, scaffolding user audits, and deriving actionable insights from user-engaged audit reports. Furthermore, practitioners shared organizational obstacles to user-engaged auditing, surfacing a complex relationship between practitioners and user auditors. Based on these findings, I discuss opportunities for future HCI research to help realize the potential (and mitigate risks) of user-engaged auditing in industry practice.

5.1 Introduction

In recent years, algorithm audits have risen to prominence as an approach to uncover biased, discriminatory, or otherwise harmful behaviors in algorithmic systems [91, 127, 168, 169, 290, 361, 417, 447, 512, 597]. Today, algorithm audits are typically conducted by small groups of experts such as industry practitioners, researchers, and activists [361, 447]. Although expert-led approaches have been highly impactful, they often suffer from major blindspots, and fail to detect critical issues. For example, expert-led audits can fail when those conducting the audit lack the relevant

cultural knowledge and lived experience to recognize and know where to look for certain kinds of harmful algorithmic behaviors [148, 256, 474, 589].

To overcome limitations of current algorithm auditing techniques, researchers in HCI and AI have begun to explore the potential of more *user-engaged* approaches to algorithm auditing, which directly engage users of AI products and services in surfacing harmful algorithmic behaviors. Recent years have seen many cases in which users organically came together to uncover and raise awareness about harmful behaviors in algorithmic systems they use day-to-day, which had eluded detection by industry teams or other expert auditors [474]. Inspired by these observations, researchers have begun to explore the design of systems that can leverage the power of everyday users and crowds to surface harmful algorithmic behaviors that might otherwise go undetected (e.g., [43, 98, 300, 315, 382, 383]). The designs of existing research systems span a spectrum of user engagement, from more practitioner-led approaches—such as crowdsourcing workflows in which users’ testing and auditing activities are more heavily guided and constrained by requesters—to more user-led approaches in which users take greater initiative in directing their own activities.

In parallel to these research efforts, several major technology companies have begun to experiment with approaches that engage users in auditing their AI products and services for problematic behaviors. For example, in 2021 Twitter introduced its first “algorithmic bias bounty” challenge to engage users in identifying harmful biases in its image cropping algorithm [114]. In another effort, Google launched the “AI Test Kitchen,” a web-based application that invites users to experiment with some of Google’s latest AI-based conversational agents, and to report any problematic behaviors they encounter [560]. More recently, inspired by Twitter’s “algorithmic bias bounty,” OpenAI initiated a “Feedback Contest” to encourage users to “provide feedback on problematic model outputs” during their interactions with ChatGPT chatbot [15].

Despite growing interest from industry, there remains a gulf between the academic research literature on user-engaged auditing and current industry practice. In particular, we still know little about industry AI practitioners’ current practices and challenges around user-engaged auditing, and what opportunities exist for them to better leverage such approaches in practice. To investigate, in this chapter we explore the following research questions:

RQ1 What are AI practitioners’ current motivations and practices around user engagement in auditing their AI products and services for problematic algorithmic behaviors?

RQ2 What opportunities and challenges do practitioners envision for user-engaged approaches to better support their algorithm auditing efforts?

We conducted a two-stage study with 12 industry practitioners from 9 technology companies, all of whom have experimented with engaging users in auditing their AI systems for problematic algorithmic behaviors. We first conducted semi-structured interviews to understand practitioners’ current practices and challenges around engaging users in AI testing and auditing. We then conducted co-design activities, working with practitioners to iteratively co-design three design artifacts as a way to further probe challenges perceived by practitioners and opportunities to better support user-engaged approaches to algorithm auditing in industry practice.

Overall, our participants shared three major motivations for engaging users in AI testing and auditing: understanding users’ subjective experiences of problematic

machine behaviors, overcoming their teams' blindspots when auditing their products and services, and gathering evidence from users to help them advocate for fairness work within their organizations. Participants shared prior experiences engaging users on different scales, from individual user study sessions, to focus group workshops, to large-scale user feedback and crowdsourcing activities. However, in doing so, practitioners encountered various challenges in engaging users effectively. For instance, practitioners discussed challenges they faced in recruiting and incentivizing the "right" group of auditors for a given task, with relevant identities and lived experiences. Participants also discussed the difficulties in scaffolding users towards productive auditing strategies, without biasing them to simply replicate industry teams' own blindspots. Finally, practitioners discussed the challenges of quantification when deriving actionable insights from user-engaged auditing reports: relying upon the majority vote runs the risk of masking the very biases an audit is intended to uncover. In addition, participants shared broader organizational obstacles to user-engaged auditing, highlighting key tensions that arise in practice when involving users in algorithm auditing efforts such as potential PR risks, profit motives that work against protecting marginalized groups, and privacy and legal concerns.

As private companies increasingly experiment with user-engaged approaches to algorithm auditing, HCI research has a critical role to play in shaping more effective and responsible practices. To this end, this work contributes:

- An in-depth understanding of industry practitioners' motivations, current practices, and challenges in effectively engaging users in testing and auditing AI products and services. Our findings shed light on the types of problems practitioners aim to address through user engagement around algorithm auditing, as well as the ways practitioners navigate organizational tensions around user involvement in AI development processes.
- A set of design implications for user-engaged algorithm auditing, beyond standard considerations for human computation or user feedback systems.
- Insights into the complex relationship between user auditors and industry practitioners working on responsible AI, suggesting opportunities for future HCI research to help realize the potential (and mitigate risks) of user-engaged auditing in industry practice.

5.2 Related Work

5.2.1 Understanding and supporting responsible AI practices in industry contexts

In recent years, significant effort has been directed towards the development of approaches, guidelines, and tools to help industry practitioners audit their AI products and services for unfair, biased, or otherwise harmful algorithmic behaviors (e.g., [14, 56, 67, 68, 368, 425, 426]). Early work in this area has largely been guided by advances in academic research on AI fairness [3, 32, 51, 172, 310, 414]. Yet in a series of interview studies and surveys with industry AI practitioners, Holstein et al. [256] found that there were major disconnects between the tools offered by the research community, versus the actual on-the-ground needs of industry AI practitioners. To address such gaps, a growing line of research in HCI has focused on better understanding industry AI practitioners needs and designing to support responsible AI

practices in industry. For example, studies from Madaio et al. [342] and Rakova et al. [419] investigated the organizational challenges and barriers that practitioners face in practice when attempting to build more responsible AI systems.

Meanwhile, to better support responsible AI practices, companies have been developing responsible AI guidelines such as People + AI guidebook [426], trustworthy AI principles [537], AI fairness checklists [425], and responsible AI toolkits such as AI Explainability 360 [14] and Fairlearn [67]. However, recent HCI research has surfaced gaps between fairness toolkits' capabilities and practitioners' needs [142, 323, 430]. For example, Kaur et al. [291] found that AI practitioners often over-trust and misuse AI explainability toolkits. Other work from Lee et al. and Deng et al. identified misalignment between the designs of existing fairness toolkits versus practitioners' actual desires and usage of these toolkits [142, 323, 430]. In interviews with AI practitioners, these authors found that, beyond the functionality provided by current toolkits, practitioners desired tools that could help them bring in perspectives from relevant domain experts and users, in order to aid them in auditing their AI systems [142]. In the next sections, we discuss emerging work that aims to support the power of users in algorithm auditing.

5.2.2 The power of users in algorithm auditing

Metaxa et al. [361] define an algorithm audit as “a method of repeatedly querying an algorithm and observing its output to draw conclusions about the algorithm's opaque inner workings and possible external impact.” A growing body of work in HCI, AI, and related communities has developed tools and processes to audit algorithmic systems for biased, discriminatory, or otherwise harmful behaviors (e.g., [91, 361, 447]). Past work in algorithm auditing has uncovered harmful algorithmic behaviors across a wide range of algorithmic systems, from search engines to hiring algorithms to computer vision applications [40, 91, 241, 379, 408, 512].

Today, algorithm audits are typically conducted by small groups of experts such as industry practitioners, researchers, activists, and government agencies [361]. However, such expert-led audits often fail to surface serious issues that everyday users of algorithmic systems are quickly able to detect once a system is deployed in the real world [256, 474]. For instance, this approach can fail when those conducting the audit lack the relevant cultural knowledge and lived experience to recognize and know where to look for certain kinds of harmful algorithmic behaviors [148, 256, 474, 589]. In addition, expert-led audits may fail to detect certain harmful algorithmic behaviors because these behaviors only arise—or are only recognized as harmful—when a system is used in particular context or in particular ways, which auditors may fail to anticipate [127, 168, 190, 256, 465, 474].

Recent years have seen many real-world cases in which *users* have uncovered and raised awareness around harmful algorithmic behaviors in systems they use day-to-day (e.g., search engines [91], online rating/review systems [168, 512], and machine translation systems [408]) although expert auditors had failed to detect these issues. Shen et al. [474] developed the concept of “everyday algorithm auditing” to describe how everyday users detect, understand, and interrogate problematic machine behaviors via their daily interactions with algorithmic systems. In the cases these authors reviewed, regular users of a wide range of algorithmic systems and platforms came together organically to hypothesize and test for potential biases. More recently, DeVos et al. [148] conducted a series of behavioral studies to better understand how users are often able to be so effective, both individually and

collectively, in surfacing harmful algorithmic behaviors that more formal or expert-led auditing approaches fail to detect. As discussed next, recent research is beginning to explore ways to support the power users in algorithm auditing to overcome limitations of expert-led approaches.

5.2.3 Supporting user-engaged algorithm auditing

Recognizing the power of users in algorithm auditing, researchers have begun to explore the design of systems to support more *user-engaged* approaches [148, 315] to algorithm auditing, which directly engage users in surfacing harmful algorithmic behaviors that might otherwise go undetected.

A line of work has developed interfaces, interactive visualizations, and crowd-sourcing pipelines to support people in actively searching for algorithmic biases and harmful behaviors [43, 98, 300, 382]. The designs of these research systems span a spectrum of user-engagement, from more practitioner-led approaches to more user-led approaches in which users take greater initiative and control in directing their efforts. For example, Ochigame and Ye developed a web-based tool called Search Atlas, which enables users to easily conduct side-by-side comparisons of the Google search results they might see if they were located in different countries to spot [383]. Kiela et al. developed a general research platform called Dynabench, which invites users to try to identify erroneous and potentially harmful behaviors in AI models [300]. Using Dynabench, users can generate test inputs to a model to try to find problematic behaviors, flag behaviors they identify, and provide brief open-text responses if they wish to offer additional context. More recently, Lam et al. developed a tool called “IndieLabel,” in order to empower end users to detect and flag problematic algorithmic behaviors and then author audit reports to communicate these to relevant decision-makers [315].

In parallel, several major technology companies have begun to experiment with approaches that engage users in auditing their AI products and services for problematic behaviors. For example, in 2021 Twitter introduced its first “algorithmic bias bounty” challenge to engage users in identifying harmful biases in its image cropping algorithm [114]. In another effort, Meta adopted the Dynabench platform described above, to discover potentially harmful behaviors in natural language processing models [300]. More recently, Google launched the “AI Test Kitchen,” a web-based application that invites users to experiment with Google’s latest LLMs-powered conversational agents, and to report any problematic behaviors they encounter, with the stated goal of engaging users in “*learning, improving, and innovating responsibly on AI together*” [560]. In addition, organizations like OpenAI and HuggingFace are beginning to include built-in interface features that invite users to report harmful algorithmic behaviors they encounter while interacting with LLM-powered applications like text-to-image generation tools. HuggingFace developed features to engage end users in flagging ethical/legal issues on their API [401]. In addition, OpenAI initiated a feedback contest around their LLM-based tool ChatGPT, with the goal of encouraging users to “provide feedback on problematic model outputs” [15].

Despite growing interest in both academia and industry, there remains a gulf between the academic research literature on user-engaged auditing and current industry practice. To date, little is known about industry AI practitioners’ current practices and challenges around user-engaged auditing, nor what opportunities exist for them to better leverage such approaches in practice. In this chapter, we take a

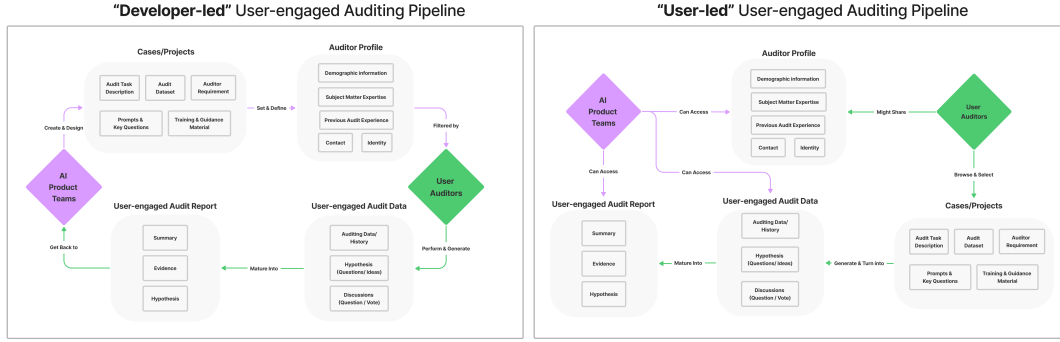


FIGURE 5.1: Two potential user-engaged auditing pipeline designs that were iteratively co-designed with participants. The left image shows the “developer-led” pipeline design, and the right image shows the “user-led” pipeline design. Each figure illustrates a possible interaction flow between user auditors and AI product teams, showing how auditing tasks are created, how background information on user auditors is shared, how user auditing reports are generated based on auditors’ findings, and how these reports are shared with AI product teams. During the co-design activity, participants could zoom in, annotate, and modify the details. We used these pipeline flowcharts as **probes**, not as **final products**, to investigate more deeply on practitioners’ challenges and desires.

first step towards understanding current practices, challenges, and design opportunities for user-engaged approaches to algorithm auditing in industry practice.

5.3 Method

5.3.1 Study design

We conducted a two-stage study involving semi-structured interviews followed by iterative co-design activities. We first conducted semi-structured interviews to understand participants’ current practices and challenges around engaging users in AI testing and auditing. In the next stage, we engaged participants in a co-design activity to further probe the opportunities and challenges in supporting user-engaged algorithm auditing in industry practice. We worked with participants to iteratively design three artifacts: a **user-engaged audit report**, representing a “wish list” of types of information that they would ideally want to solicit through a user-engaged auditing approach, and **two user-engaged auditing pipelines**, building upon initial designs informed by interview findings and insights from prior literature [148, 361, 447, 474]. Throughout the study, we iterated on these design artifacts based on feedback and design ideas from prior participants. We used these artifacts and the *process* of co-designing them to provoke deeper conversations around participants’ desires, as well as potential risks they anticipate, for new systems that support user-engaged auditing.

Following an iterative co-design process similar to prior work (cf. [255, 343]), for our first five participants, we ran the two stages of our study in separate sessions in order to better inform the design of the initial versions of the artifacts based on the needs and desires these participants expressed in the first set of interviews. However, we soon encountered difficulty in retaining industry participants due to their busy schedules (e.g., one participant was not able to return to complete the co-design). Therefore, after our first five participants, we ran both stages in a single session. We then continued to iterate on the artifacts during the study sessions themselves. Below, we describe each of these activities in more detail.¹

¹We also provide our interview and co-design protocol in the supplementary material.

Stage one: Semi-structured interviews

To understand practitioners' current practices around engaging users in AI testing and auditing, we conducted semi-structured interviews, each lasting up to an hour. We adopted a directed storytelling approach [174]. We first asked participants to describe their team's prior experiences in trying to detect or address biased or harmful behaviors in their AI products or services, with a specific focus on whether, why, and how they engaged users in the process. For example, we asked "*Could you describe how your team attempted to engage users in auditing the AI products and services you mentioned?*" and "*What motivated you or your team to engage users in this way?*" Through follow-up questions, we probed deeper into challenges participants had encountered when attempting to engage users in the auditing process. As participants shared specific challenges they had encountered, we also invited them to share ideas for potential solutions to these challenges. For example, in response to specific challenges raised by participants, we asked "*How did your team attempt to tackle these challenges?*" and "*How effective were your team's approaches?*"

Stage two: Iterative co-design activities

To further envision future opportunities and solicit potential challenges and risks for user-engaged algorithm auditing approaches, following the interviews, we then involved participants in a series of co-design activities, following an iterative co-design process similar to prior work (cf. [255, 343]). This stage of the study lasted up to 45 minutes, and involved participants in co-design around three design artifacts: a **user-engaged audit report** and two **user-engaged audit pipeline flowcharts**. We first designed initial versions of these artifacts based on participant needs and desires expressed during stage one, as well as prior research on user-engaged algorithm auditing [148, 474]. We then iterated on their designs with practitioners throughout the co-design activities. We note that these design artifacts were not the goal of our study, but rather served as tools to probe more deeply on practitioners' challenges and desires. Below, we describe the process of co-designing these three artifacts, and how we used this process to probe on future opportunities and risks of user-engaged audits.

User-engaged audit report: We invited each participant to contribute to the design of a report that they would ideally like to see as the *output* of a user-engaged auditing process. We first asked participants open-ended questions such as "*What information would you ideally want the service to report back to your team?*" and encouraged them to sketch as they generated new ideas. To help participants come up with ideas, we presented participants with example of actual written responses generated by users during a user-engaged algorithm auditing workshop in prior work² [148]. Viewing examples of actual user responses on an auditing task provided an opportunity for participants to reflect upon gaps in their own report designs, and to anticipate potential challenges in soliciting useful information from user auditors (cf. [254, 343]). We then presented participants with a user-engaged audit report template that we initially designed based on previous work and iterated through the design during the study. Our goal was to further probe participants' feedback by providing them with potential content that an audit report can consist of (such as information about the auditors, details of the reported issue, evidence, the severity of the issue, etc.). However, we intentionally showed this report after participants generated ideas about what a report can include to avoid biasing them towards a

²We provide the example user report we used in the study in the supplementary material

	Company size	Job title	Types of AI products and services	Experience with user-engaged auditing	AI fairness experience	Fairness work part of official role?
P1	1000 - 4,999	Director and Product Lead	AI-powered knowledge graph for academic literature search	Engaging search users in assessing potential biases in the underlying knowledge graph.	4 years	Yes
P2	10 - 50	Senior Director	ML model to predict potential donors	Engaging marginalized community members in surfacing potential biases in their team's ML model	3 years	No (self motivated)
P3	25,000+	ML Engineer	Natural language processing (NLP) applications	Engaging users in rating the risk of representational harms	2 years	Yes
P4	25,000+	Senior Technical Lead	A diverse range of AI products built by their customers	Engaging users in auditing a range of AI products built by their customers	3 years	Yes
P5	25,000+	Senior Product Manager	Sentiment analysis; OCR recognition	Leading several groups of AI product teams on engaging users in testing and auditing their AI products	5 years	Yes
P6	25,000+	UX Researcher	NLP-powered products (e.g., conversational agents)	Engaging users in testing conversational agents for potentially harmful behaviors	2 years	No (self motivated)
P7	25,000+	UX Researcher	Computer vision applications (e.g., image search)	Building an interface to engage users in auditing their image search engine	2 years	Yes
P8	25,000+	ML Researcher	Information retrieval and image processing	Building internal crowdsourcing tools for AI auditing	3 years	Yes
P9	5,000 - 24,999	Data Scientist	Recommendation system	Engaging users and impacted stakeholders in auditing their recommendation system	3 years	Yes
P10	25,000+	UX Researcher	Large language model-based conversational agent and generative image model	Building a web-based application to engage users in flagging the potential biased and harmful behavior in a conversational agent	2 years	Yes
P11	5,000 - 24,999	Senior Researcher	Recommendation system	Leading research efforts and producing concrete organizational policy around user-engaged algorithm auditing	2 years	Yes
P12	25,000+	UX Researcher	NLP	Engaging marginalized communities in auditing biases in their NLP products for low-resource languages	3 years	No (self motivated)

Table 5.1: Summary of participants' backgrounds and relevant experience.

specific report format, yet giving them the opportunity to discuss other options and iterating through the report design based on their initial ideas.

User-engaged audit pipeline flowcharts: We also co-designed two opposing caricatures of user-engaged audit pipeline designs with participants which varied in the degree of initiative users assumed in the auditing process: (1) a **“developer-led”** pipeline, in which the audits were primarily initiated and coordinated by the developers of an AI product or service; (2) a **“user-led”** pipeline, in which the audits were primarily initiated and coordinated by users (see Fig. 5.1). For the developer-led pipeline, the AI product teams can fully control what AI systems (or what parts of the AI systems) should be audited, who should be considered eligible auditors, and how the auditing should be executed. In the “user-led” pipeline, users are those who initiate and control the auditing based on their interactions with the AI systems. Users could collectively initiate the auditor selection criteria, defining auditing protocol, generating audit data, and synthesizing reports. In this case, AI practitioners can only access the audit data without getting involved in or having a say in the auditing process. While we anticipated that *neither* of these caricatures would represent ideal designs from industry practitioners' perspectives, we presented these in order to *provoke further discussion* about design trade-offs between greater user versus developer control in auditing processes.

5.3.2 Participants

We adopted a purposive sampling approach [102], with the aim of recruiting industry practitioners who either (1) had direct prior experience employing *user-engaged algorithm auditing* approaches, or (2) had an interest in such approaches and had adjacent experience *crowdsourcing* approaches as part of their AI work. Specifically, using an online screening survey, we recruited members of industry teams that design and

build AI products and services, and who had *already* attempted to engage users in detecting fairness-related issues in their AI systems. In addition, we opened up the study to interested practitioners who had *not* yet experimented with user-engaged approaches to algorithm auditing, but who had prior experience using crowdsourcing approaches in other areas of their AI work. We broadened our criteria to include these participants because we expected that prior experience with crowdsourcing would help participants envision ways user-engaged approaches might support their algorithm auditing efforts. In the end, however, *all* 12 of our participants had direct prior experience experimenting with user-engaged approaches to algorithm auditing (see Table 5.2). In addition, all but three of our participants had prior experience with crowdsourcing methods.

We recruited our participants through social media (e.g., LinkedIn and Twitter), and through direct contacts at large technology companies. As discussed in prior literature that studies responsible AI practices in industry (e.g., [142, 256, 323]), recruitment for such studies can be highly challenging. Practitioners are often wary of participating in such interview studies, for instance, given that participation may require admitting the existence of flaws in their products and services that have not been made public, or sharing disagreements about their companies' current organizational culture. Although we assured potential participants that their responses would be carefully de-identified, as discussed below, we expect that such concerns likely had an influence on our recruitment.

In total, 25 practitioners completed the recruitment screening form, of which 18 met our recruitment criteria. Ultimately, 12 of these practitioners, spanning 9 companies, responded to our study invitation and participated in the study. All 12 participants participated in the interview session; all except one participant participated in the co-design session. (P4 was not able to return to complete the co-design activity due to busy schedule.) All participants were compensated at a rate of \$35 per hour for their participation. Table 5.1 overviews participants' job titles, their years of experience with user-engaged auditing, their company size, the types of AI products or services they worked on, and their experiences with user-engaged auditing. While 9 of our participants conducted user-engaged algorithm auditing as part of their main job function; three participants (P2, P6, P12) engaged users in algorithm auditing on their own initiative, to help them in advocating for fairness issues to be addressed within their organizations.

Following prior work on responsible AI practices in industry [142, 256, 342, 343], to avoid identifying individual participants who work at the forefront of sensitive topics, details about participants' demographics are omitted, and we abstract some details about participants' companies and roles. In addition, we assured participants that we would not ask them to reveal any confidential or personally identifying information about their colleagues and that we would de-identify all responses at the individual, team, and organization levels. Finally, participants were instructed that they were free to skip any questions they were uncomfortable answering, or to leave the session at any time for any reason.

5.3.3 Data analysis

Our study sessions yielded approximately 15 hours of audio that we transcribed. To analyze our interview and co-design session transcripts, we adopted a reflexive thematic analysis approach [80]. Two of the authors met after each interview and co-design session to conduct an interpretation session, and then conducted open coding of the transcripts. Throughout this coding process, the authors continuously

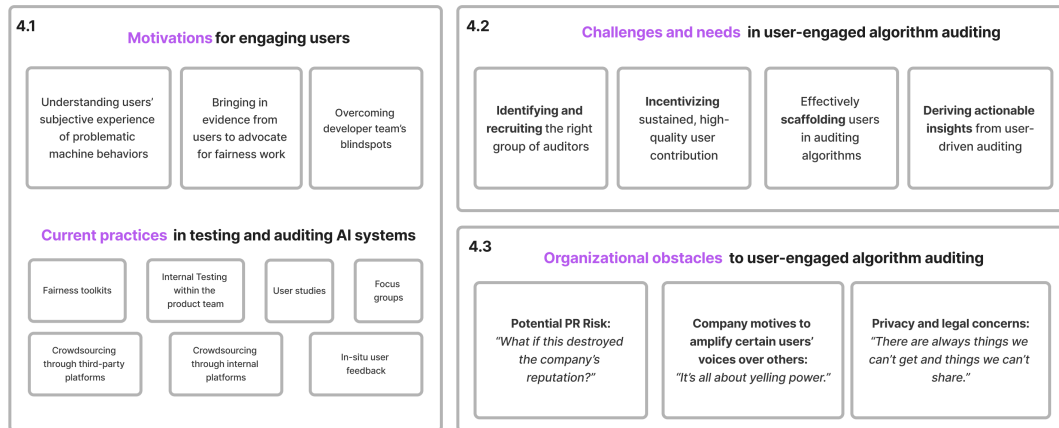


FIGURE 5.2: High-level overview of our findings. We first share participants' existing motivations and practices for user engagement around algorithm testing and auditing (Section 5.4.1). We then describe challenges participants have faced in their attempts to effectively engage users (Section 5.4.2). Finally, we share broader organizational obstacles to user-engagement around algorithm auditing perceived by participants, highlighting key tensions that arise in practice when involving users in algorithm auditing efforts (Section 5.4.3).

discussed discrepancies in interpretation, and iteratively refined the codes based on these discussions [80, 356]. In total, we generated around 1,125 unique codes. Through an iterative, bottom-up affinity diagramming process, we grouped codes into successively higher-level themes. The first level clustered our 1,125 codes into 271 themes. These were then clustered into 59 second-level themes, 15 third-level themes, and three final themes. We present our results in the following section, organized around our three final top-level themes.

5.4 Findings

In this section, we describe how industry practitioners navigate the complicated process of engaging users in surfacing harmful algorithmic behaviors. As discussed below, practitioners mediate conflicts between (1) the underlying values of user engagement in algorithm testing and auditing, (2) inherent challenges in effective user engagement, and (3) the cultural, legal, and organizational obstacles that disincentivize bringing users' voices into algorithm testing and auditing processes. We show that practitioners see clear advantages of engaging users in and around algorithm auditing processes, which have led them to explore leveraging crowdsourcing platforms and the design of new interfaces that enable in-situ feedback from their users. Yet practitioners also discussed complexities of engaging users in surfacing algorithmic harms, which introduce unique challenges beyond those faced with conventional human computation or user feedback systems. We summarize our findings in Figure 5.2.

5.4.1 Practitioners' motivations and practices for engaging users to audit AI products and services

Participants believed that engaging users in testing and auditing AI products and services could help them understand users' subjective experiences of problematic machine behaviors and help to overcome developer teams' blindspots. Participants also noted that having direct reports of potential issues from users could serve as

powerful ammunition when making the case internally for a particular course of action. Driven by these motivations, participants reported having experimented with user engagement at various scales to test and audit their systems for problematic behaviors.

Motivations to engage users in auditing AI products and services

In our semi-structured interviews, all 12 participants reported that they currently use fairness toolkits or internal auditing with the developer teams to detect problematic behaviors in their AI products. However, participants quickly encountered limitations with these approaches, which prompted them to explore user-engaged approaches as an alternative. Below, we discuss three major motivations that participants shared for experimenting with more user-engaged approaches to algorithm auditing.

Understanding users' subjective experiences of problematic machine behaviors: Participants were motivated to adopt user-engaged approaches because in many settings, they felt that it was not possible to measure “unfairness” or “harmfulness” without an understanding of users' perceptions. As shown in Table 5.2, eight of our 12 participants had previously used open-source ML fairness toolkits (e.g., Fairlearn [67], AIF360 [13]) to attempt to audit their teams' AI systems. Participants said these toolkits only offered aggregate fairness metrics, whereas appropriately assessing algorithmic biases and harms in their systems required in-depth, qualitative assessments (e.g., for representational harms [129, 256]). For example, P3 shared that their main role initially involved quantitatively assessing potential biases in NLP applications using fairness toolkits. However, P3's team soon sought out user voices, because *“things like representational harms were difficult to quantify and analyze using fairness toolkits, without understanding the reason behind why users with certain identit[ies] feel uncomfortable or offended.”* To this end, P3's team conducted focus groups with users who encountered harmful biases in their NLP applications, in order to better understand users' *experiences* in their own words.

Several participants described sometimes needing help from users to *determine* what “fair” or “non-harmful” behavior would look like in the first place. For example, P9 said a common approach when auditing a recommendation system is to assess *“how diverse the recommendations on your platform are.”* However, different users often have conflicting perspectives regarding what it means to have appropriately “diverse” recommendations. P9 believed the way to understand this was to *“directly ask them[users] what they think.”* Similarly, P11 emphasized that in order to empower their team to design effective remediations, *“It's not enough to just have people flag that a recommendation is stereotypical, we want to understand why they think it is so that our team could brainstorm potential solutions.”*

Overcoming developer teams' blindspots: Another motivation for adopting user-engaged testing and auditing approaches was to overcome cultural and experiential blindspots among product team members. Ten of our 12 participants said they conducted internal testing within the product team (see Figure 5.2, column 2). However, echoing findings from Holstein et al. [256], participants reported that internal testing with small groups of developers often resulted in blindspots, prompting them to involve users to surface *“unknown unknowns”* (P10), or the issues that the team does not know exist. For example, P8 noted that the images on their platform came from all over the world, involving signs and languages that were deeply cultural and regional, so *“it would never be possible for [their] team to capture all these diverse aspects.”* Similarly, P5 complained that *“a lot of [their] AI ethics and bias activities*

	Approaches that do not engage users		User-engaged approaches for testing and auditing their AI systems				
	Fairness toolkits	Internal testing within the product team	User studies	Focus groups	Crowdsourcing through third-party platforms	Crowdsourcing through internal platforms	In-situ user feedback
P1	X	X		X		X	X
P2		X	X				
P3	X	X		X	X		
P4	X		X	X			
P5	X	X	X	X	X	X	
P6		X	X	X			
P7		X	X		X		X
P8	X			X		X	X
P9	X	X	X			X	
P10	X	X	X	X			X
P11	X	X	X		X	X	
P12	X	X		X		X	

TABLE 5.2: Participants’ existing practices for testing and auditing their AI systems for harmful behaviors and biases. In this table, we report both approaches that do not engage users and user-engaged approaches. Section 5.4.1 describes the *limitations* of approaches that do not engage users. Section 5.4.1 describes the user-engaged approaches in detail. In this table, approaches that do not engage users are color-coded in gray; approaches for user-engagement at small scales are shown in light green; and approaches for user-engagement at larger scales are shown in darker green.

only contain [their] own employees, and the perspectives are extremely limited.” P5 found that bringing in perspectives from external stakeholders helped to surface problematic algorithmic behaviors that their internal teams had never considered.

Some participants specifically emphasized the importance of engaging marginalized community members in auditing their AI products. P12 said, “*No one in the developing team speaks the language and knows the idioms — how would they properly audit the outcomes? That’s why I have been spending time bringing native speakers into the auditing process.*” When P11’s team first started auditing the recommendation system in their product, they attempted to use personas to simulate real-world stakeholders and bring empathy to the team members, but they soon realized their blindspots persisted, as “*the hypothetical personas are still basically what we imagined based on our understanding [...] it’s just not realistic for a group of white people to truly understand Black artists’ perspectives through fictional cards.*”

Bringing in evidence from users to advocate for fairness work: Finally, five participants shared that one of their major motivations for adopting user-engaged approaches was to gather evidence, in the form of direct quotes from users, that could help them persuade others on their teams that an issue was worth addressing. This was a particularly important motivation among three participants (P2, P6, P12) who were *self-motivated* to address fairness issues in their teams’ products and not in roles that directly incentivized and supported this work. For example, P2 shared that, while building AI services to predict potential donors to their client institution, conversations with donors from marginalized communities inspired their team to “*fundamentally reevaluate the potential biases towards who will donate in their dataset.*” As another example, P6 said their “*engineers were pretty confident about the performance of [their] model and just making assumptions about the real-world situation when people are using the tool actively,*” rarely communicating with UX researchers like P6. However, P6 “*got the luxury to chat directly with folks who are working on building the product*” when they brought in users’ voices and activities about the tool. Similarly, P4 shared that when data scientists voice concerns about the ML model being unfair and biased using numbers and graphs, “*it is not a tangible risk to a business owner [...] bad experiences and feedback from the users work way better than numbers to motivate business owners to think deeper about the product’s potential negative impact.*”

Existing algorithm auditing approaches that engage with users

Driven by the motivations discussed above, all 12 of our participants reported having experimented with user-engaged approaches to test and audit their systems for problematic algorithmic behaviors. As shown in Table 5.2, these methods varied in scale: practitioners conducted user studies and focus groups with small groups of users, but they also attempted to engage a larger number of users in the auditing process through crowdsourcing tools or in-situ user feedback.

User studies and focus groups: Nine participants shared that they have conducted single-person user studies, and eight said they have conducted focus groups, to engage users in testing their AI services and products. Participants shared how working with users provided them with different and new perspectives on a product and its potential biases. To expand their perspectives, some practitioners specifically sought out users who belonged to marginalized groups. For example, in order to mitigate potential biases in their donor database—which was used as the training data set for their AI service to identify potential future donors—P2’s team connected with prior donors who have minority backgrounds and were *“not being included in the current database [that was] full of rich white people”* in order to understand their donating perspectives and experiences. This approach helped P2’s team realize that their database prioritizes privileged race and gender groups and change their overall problem formulation and development strategy for their models. Similarly, P12 helped their machine translation team better understand how errors in their products impacted marginalized immigrant communities through *“chatting with immigrants living in the US with low language proficiency both individually and in groups.”*

Crowdsourcing and in-situ user feedback approaches: Although talking to users provided practitioners with new perspectives on potential harms their products might introduce, practitioners desired more scalable approaches to reach larger and more diverse groups. As a result, 9 of our 12 participants reported that they had previously leveraged *crowdsourcing* or *in-situ user feedback* approaches to attempt to engage users in surfacing harmful machine behaviors *at scale*. This included using both third-party and internally-built crowdsourcing tools, as well as soliciting in-situ feedback directly from users during everyday interactions with their products and services.

In our semi-structured interviews, some participants reported using third-party crowdsourcing platforms like Amazon’s Mechanical Turk or company-internal crowdsourcing platforms to invite a large number of crowd workers to help audit their AI products. For example, P3’s team found it useful to conduct focus group workshops to have users audit their NLP applications. However, given the number of languages that their applications cover, their team turned to more scalable methods. For instance, to audit a model intended to detect offensive sentences, P3 deployed a task on Amazon’s Mechanical Turk to *“ask crowd workers if they find the sentence offensive or not, and ask them to flag the offensive ones.”* This approach helped P3’s team uncover a large number of offensive sentences that were previously not predicted as offensive by their language models. Similarly, P5’s team paid crowdworkers on a third-party crowdsourcing platform to detect potential biases exhibited in their optical character recognition (OCR) model towards hand-written characters in different source languages.

Inspired by recent cases in which users surfaced biases through their day-to-day interactions with the AI systems (e.g., [114, 474]), four participants developed features that were *seamlessly integrated into their AI systems’ user interfaces*, to allow users to provide feedback on potential biases and harmful behaviors in-situ, during their

everyday interactions with a system. For example, P7 led an effort to build a plug-in to help users report any harmful outputs encountered while using their image recognition service. Similarly, P10 led an effort to build a web-based application where users could interact with an unreleased conversational agent prototype and report any problematic algorithmic behaviors they encountered during their interactions with the prototype.

5.4.2 Challenges and needs in user-engaged algorithm auditing

Participants noted that although their current crowdsourcing and in-situ user feedback approaches could mitigate some of the limitations of existing auditing approaches, they still faced numerous challenges in effectively engaging users beyond existing design considerations for human computation or user feedback systems. For example, detecting harmful algorithmic behaviors requires recruiting and incentivizing the right group of auditors, with relevant cultural backgrounds, lived experiences, and perspectives. In addition, designing auditing tasks can introduce unique complexities, beyond those faced in conventional human computation tasks: the tasks need to guide users towards productive auditing strategies, but without overly influencing them to simply replicate industry practitioners' own biases and blind spots.

Identifying and recruiting the right group of auditors

All 12 participants emphasized the importance of identifying and recruiting users with relevant identities, cultural backgrounds, and expertise to better test and audit the AI systems.

Identifying relevant subgroups of users: While discussing the design of the “developer-led” user-engaged auditing pipeline, many participants (N=6) shared that they found it challenging to determine which demographic subgroups were most critical to engage in auditing specific AI products. For example, P6’s team wanted to engage real-world users to audit their conversational AI products, yet they did not know *“who are these users, how much they will be impacted by our product, and how to reach out to them.”* When P7’s team attempted to recruit users to assist in auditing their image search service, they became overwhelmed as they *“started to think about the intersectionalities [of users], and the demographics just blew up into a billion different categories”* leading P7’s team to wonder *“what is the right level of identity intersections to look at?”* P7 noted that this challenge caused their team to *“get lost”* in the process of setting up their own user-engaged auditing pipeline. Similarly, P9, whose team worked on building fairer recommendation systems, shared, *“We were having [a] hard time defining the genre of the content from the artists, and we don’t want to just label the artists by ourselves and project our biases [towards] the users even before we started to engage them.”* Given this experience, P9 viewed the challenge of identifying relevant subgroups of users as *“a fundamental but intractable first step”* to conducting user-engaged auditing on their product.

Recruiting a diverse and representative set of users: All 12 participants shared that, even when they knew what identities or background expertise they wanted to target, their current approaches were inadequate to actually recruit the targeted groups of users. Six participants reported that, in order to recruit user auditors, they relied heavily on personal networks or existing relationships with previous users. As such, they encountered challenges in recruiting users auditors from demographics, domains, regions and cultures they had not established such relationships with. For

example, P1 shared that, as a US-based company with few Asian employees, their team encountered challenges in recruiting *“users from Japan and Korea”* to judge, for example, whether their AI services generate *“potentially offensive labels for a sea area between Japan and Korea, given that these two countries had previous conflicts on the naming and jurisdiction of that area.”* Similarly, P12 mainly relied on their team members’ existing personal relationships with customers to recruit marginalized community members, but often stumbled when attempting to reach out to a community that none of their team members had previously interacted with.

During the co-design sessions, several participants (N=9) expressed the belief that future tools that support more *“user-led”* approaches to auditing, which appeal to users’ *intrinsic* and *social* motivations to participate, could be helpful in reaching users with specific identity characteristics and/or domain knowledge. For example, P5 drew an analogy to Wikipedia, believing that a more *“user-led”* process could attract users to voluntarily and collaboratively audit AI systems that affect their lives, similar to how *“people collectively edit articles on Wikipedia based on their interests, and will attract people with similar interests to join.”* Similarly, P12 saw potential for more *“user-led”* approaches to user-engaged algorithm auditing to organically attract people with similar identities and shared experience: *“For example, people on Reddit with shared identities will come together and discuss problems they are facing.”*

Incentivizing sustained, high-quality user contributions

While previous work has shown both external and intrinsic motivators driving online collective actions [260, 309, 329], in our study, we found that practitioners currently rely primarily upon external motivators, such as financial and social incentives, to motivate user auditors to make sustained, high-quality contributions. Meanwhile, participants shared challenges in employing these motivators.

Challenges in implementing financial incentivization in more *“user-led”* pipeline:

All 12 of our participants shared that they currently provide financial compensation to motivate user auditors. While participants found it relatively straightforward to implement financial incentivization in a more *“developer-led”* auditing pipeline, during the co-design activities, five participants raised challenges around how to compensate users in a *“user-led”* auditing pipeline that was more exploratory and discussion-based. For example, P7 commented that the *“open-endedness”* in the *“user-led”* auditing pipeline also made it difficult to decide how to compensate user auditors: *“How do you pay people in this context? Right? Like what with the task-based things, it’s like, there’s a clear incentive: you do the tasks, you get paid. Here, where it’s exploratory, [do] you pay people for just spending time in this interface? Do you pay them for just chatting? Or just generating hypotheses?”* Similarly, P10 shared a prior experience where their team *“had discussed paying the users who gave good amounts [of] and quality feedback”*, but the conversation died when their team struggled to come up with clear definitions of *“good amounts”* and *“quality”* to implement a concrete compensation plan.

Benefits and risks of social motivators: P1’s team leveraged social motivators like *“peer recognition”* and *“social interaction”* by creating a *“star system”* to reward high quality auditors, as they found users enjoyed earning stars to demonstrate their *“high reputation.”* P1’s team further *“implemented a leader board to keep track of who is bringing in the best feedback, the most feedback, and whose feedback is being endorsed by*

lots of other auditors,” and to allow users to write “*recommendation letters*” for one another to audit other tasks. During the co-design study, several participants (N=8) believed that a more “user-led” auditing pipeline could potentially amplify social motivators. Nevertheless, participants raised concerns that “*certain user groups’ voices might be further marginalized in [the ‘user-led’] pipeline*” (P9). To combat this, participants suggested developers should intervene and facilitate the conversations among the user auditors, to amplify marginalized voices throughout user-engaged auditing processes.

Effectively scaffolding users in auditing algorithms

Participants noted that it was challenging, in practice, to design user auditing tasks and instructions that could empower users to generate meaningful insights about their AI products and services. As we discuss below, participants shared several challenges they had faced in guiding user auditors without imposing the development team’s own biases upon them, and in prompting user auditors to provide more critical feedback on an AI system’s overall design.

Guiding users towards productive auditing strategies, without overly biasing them: During the interview portion of our study, participants shared experiences where user auditors had misunderstood the tasks they were given or had failed to provide sufficient detail and context for industry teams to act on their reports. Thus, throughout the co-design portion of our study, several participants (N=8) expressed desires for ways to help user auditors better understand their team’s intended goals for an audit, and to scaffold them in auditing a system more effectively. For instance, P5 suggested “*sharing with the users a theoretical structure of biases, an algorithmic harm taxonomy*” to reference both during an onboarding phase and at any point during their auditing activities. However, P5 was uncertain what such a taxonomy would look like in their context (i.e., sentiment analysis). Similarly, P8 and P10 desired better ways to nudge user auditors to “*think out of the box*” (P8) and “*ask hard questions [to a conversational agent] and break the model and surface our ‘unknown unknown’*” (P10). P7 noted that, in the context of image search, user auditors may not always test the impacts of small perturbations, so they suggested prompting users to do so: “*Chang[ing] a small word might lead to a very different search result, and we definitely want to guide users to explore these small changes*” (P7).

Despite this desire to guide user auditors in more productive directions, some participants (N=6) expressed concerns that providing too much or the wrong kinds of scaffolding might bias users to think too much like their own teams—potentially limiting the value of a user-engaged auditing approach. For example, during our co-design activity, P10 noted that designing guidelines and specific prompts for user auditors was “*quite tricky since of course we want to offer detailed guidelines and ask specific questions like ‘do you think this output is biased towards Asians or women,’ but our questions might actually bias the users when they are finding biases. [...] We need to make sure that we don’t let our confirmation bias affect this [user-engaged audit process]*”.

Acknowledging the challenge of navigating these tradeoffs, five of our participants emphasized the importance of cross-functional collaboration in designing effective guidelines, tasks, and prompts for user auditors. For example, in most of these teams, the design of auditing tasks was left to engineers who had no training in the design of human subjects research methods. By contrast, P10 worked on a team where UX practitioners were involved in the design of auditing tasks. However, the UX and AI teams often worked in silos. P10 felt that “*with a UX background,*

I only have a surface understanding of large language model[s]. I can't design a good auditing task just by myself if we eventually want to incorporate some auditing feedback from users into the current model."

Soliciting critical and holistic feedback from user auditors: Ten of our 12 participants expressed desires for better ways to prompt more critical and holistic feedback from user auditors. While iterating on the user-engaged audit report and the "developer-led" auditing pipeline, P9 suggested that in order to prompt critical feedback, it could be helpful to share the team's *rationales* behind particular design decisions, which might otherwise remain opaque to user auditors. P9 shared a prior experience where their team had initially struggled to gather the sorts of critical feedback they were hoping for. However, after their team shared more details about specific design rationales, "[user auditors] asked questions like, 'Do I actually want to get recommendations in this way? What's a better way to design this that fits my preference?' These types of questions are the ones we wish our users [would] ask when auditing and include [in] their final report." Similarly, while describing their team's prior experiences with user-engaged audits, P7 mentioned that users sometimes express a desire to know more about "*why and how*" their AI products and services are designed, in addition to seeing the AI's outputs, so that they could surface potential *procedural* issues that might not otherwise be as visible to them.

Several participants believed that a more "user-led" algorithm auditing process could help to catalyze more critical inquiries from users. For example, while iterating on the "user-led" audit pipeline, P12 said, "*when it's 'developer-led,' we are still testing if there are mistakes or unfairness in the product, right? But what if users believe this AI product shouldn't even exist? This is something you can get by giving people more freedom to discuss.*" For this reason, P1 believed it was critical to "*allow [user auditors] to chat with each other,*" in contrast to conventional crowdsourcing approaches, in which crowdworkers perform tasks in independent silos. Similarly, P10 argued that platforms or tools for user-engaged algorithm auditing should include mechanisms "*for users to share and discuss the issue they found during the auditing process*"

Deriving actionable insights from user-engaged auditing

Seven out of 12 participants shared challenges they currently face in deriving actionable insights from the user audit reports. In particular, participants shared that, unlike in conventional crowdsourcing approaches, understanding the perspectives of "outliers" may often be more important than understanding the majority view. In addition, participants found it challenging to communicate qualitative auditing results to key decision-makers within their organizations, given a culture of valuing numbers over more complex stories.

"It is no longer simply checking the majority vote." Aggregating and interpreting user-engaged auditing reports: Throughout our co-design activities, all 12 participants highlighted the challenges of aggregating and interpreting results from user-engaged auditing processes. As P3 put it, "*in more traditional numerical crowdsourcing activities, you would throw away a person who like always contradicts what everyone else says.*" However, P3 noted that the "outliers" in a user-engaged audit are often the ones that developer teams care the most about. These "outliers" may represent users in the margins, who are sensitized to issues that other auditors are not: "*So maybe like everyone said, sample A was not offensive or problematic, except for like, auditor number 39 [...] because number 39 actually found problematic things others didn't*" (P3). Similarly, P2 said that in their view, "*a few [user] audit results stating potential biases and harms might weigh more than one hundred similar good audit results*", similar

to doing UX tests when *“a single negative review might surface key insights for room to improve.”*

The challenge of interpreting results in aggregate becomes especially hard when the number of user audit reports gets large. P7 stressed the importance of collecting *“‘why’ information”* from users, such as open-text responses explaining why a user perceives a particular algorithmic behavior as problematic. However, P7 complained that they currently lack an *“efficient mechanism to combine quantitative and qualitative insights from users’ feedback [...] it is no longer simply checking the majority vote.”* While discussing the user-engaged audit report, P5 said, *“When I have thousands of training data annotations from crowd workers, I could just check the statistics of the aggregated results [...] but now with thousands of these audit reports, how am I supposed to figure out the most valuable information?”* P5 believed that in order to institutionalize user-engaged auditing in their organization, their team needed to invest in developing new automatic pipelines to augment their current manual process of reviewing user-engaged audit reports.

“It’s just our current culture, we still believe more in the numbers”: Quantification and its challenges: All 12 of our participants mentioned that, in order to effectively integrate findings from a user-engaged audit, they need to be able to present clear, quantifiable metrics to leadership and other team members. As P11 put it, *“It’s just our current culture; we still believe more in the numbers.”* However, this often presented challenges in the context of user-engaged algorithm auditing. For example, when P12 could not offer a clean cut number but only *“a complex story”* in response to questions from developers such as *“[what] percentage of users believe their recommendation is bad,”* they were told by their product manager that they were being *“distracting and counterproductive”* to project progress. While iterating on the user-engaged auditing report, P6 shared related concerns about how to measure the progress of a user-engaged audit: *“One consideration for combining these user-engaged reports is, what is the metric to define ‘success’ here? How many [reports] is enough? How much more do we need before we stop?”* P6 suggested that in order to effectively translate these reports into concrete actions from the product team, defining such metrics and clearly scoping the goal of a user-engaged audit would be critical given that *“[monetary and time] cost is [always a] concern.”*

5.4.3 Organizational obstacles to user-engaged algorithm auditing

Beyond challenges in effectively engaging users, participants also shared broader, organizational obstacles they perceived around potential PR risks, profit motives that work against protecting marginalized groups, and privacy and legal concerns. Taken together, our findings shed light on the ways practitioners currently navigate organizational tensions specifically around user-engagement in algorithm testing and auditing.

“What if this destroyed the company’s reputation?” Potential PR risk

Multiple participants (N=6) raised concerns regarding the feasibility of full institutional buy-in to user-engaged auditing from their organizations. Participants were especially skeptical of the more “user-led” pipeline, as it seemed to hold the greatest potential PR risk. For example, while co-designing the “user-led” pipeline, P7 shared, *“PR issue[s] are...] one main reason I see companies don’t want [user-led algorithm auditing] as an everyday thing.”* P7 backed this fear with a specific experience: *“We had users just try to find the absolute worst thing possible in our models and [they] made it into*

a story for social media instead of reporting back to us.” P7 also said their team leadership worried that this behavior could “expose the vulnerability of [their] models to their competing companies” and that ultimately “involving users might create more headlines that damage the company’s public image”. Similarly, P3 said these fears constituted “a major reason why [their] company is still experimenting [with user-engaged approaches] on some applications instead of making it a company-wide thing.” Even P8, whose company began user-engaged auditing after PR pressure, worried that user-engaged auditing could cause new PR issues. After co-designing the report, P8 said, “This report would be extremely useful if it only goes to us,” and wondered how to “hold the users accountable and make sure they don’t destroy the company’s reputation after gaining the trust.”

“It’s all about yelling power”: Company motives to amplify certain users’ voices over others

An important goal of user-engaged algorithm auditing is to translate the problems users find into concrete remediations from the product teams. However, eight of our 12 participants mentioned that, realistically, companies will prioritize addressing issues raised by certain groups of users over others. For instance, P1 said, “If [...] a large group of researchers from [a major US-based research institute] and a single user from a community college both raised concerns about our knowledge graph, unfortunately our business team would have to prioritize the former.” P1 concluded that when addressing issues raised in user-engaged audits, “It’s all about [the] yelling power of the users.” During the co-design activity, P12 stated that in order to incorporate user-engaged audit in their day-to-day AI work, “the biggest challenge is not to design the perfect workflow, but to make sure [the company] wants to do [user-engaged auditing] for social good, [not just] for earning more money from more people.” Similarly, P4 shared that in their current user-engaged auditing work, they constantly find themselves battling the business teams since “the business teams sometimes choose to neglect the users’ audit outcome if the reports were not from their ‘original [target] audience.’” P4 believed that, if their team is not implementing remediation in response to reports from the most marginalized users, their organizations “run the risk of participating in an ethics-washing activity.”

Several participants also shared that they often ran up against a “vicious cycle” (P12), in which a dearth of data from low-resource areas and marginalized communities makes it difficult for practitioners to advocate for more resources to address these areas. P12 said they struggled to get enough resources to test their language technologies with marginalized communities, as their data scientists required large-scale, quantitative evidence before approving studies with new groups of community members. Yet as P12 noted, “There is, of course, not enough evidence, when these groups were not even considered as users in the first place.”

“There are always things we can’t get and things we can’t share”: Privacy and legal concerns

As mentioned in Section 4.2.1, participants highlighted that access to user auditors’ demographics and other background information is critical in order to assign tasks to appropriate user auditors and to understand which perspectives are represented in reports from user auditors. However, all 12 participants also shared that it was often challenging to obtain certain demographic information due to privacy and legal concerns—mirroring challenges that practitioners face in AI auditing work more broadly [109, 256, 539]. However, beyond standard concerns around the collection and use of sensitive data for AI auditing, participants also shared concerns about

the data user auditors might share on an auditing platform, as well as the data they might need to share with user auditors to enable effective audits. For example, when discussing the “user-led” auditing pipeline, P7 noted the challenges that could arise if a user auditor were to “[take] somebody else’s photo and share it with other users for the purpose of auditing” or “[share] any users’ race and gender with other users if they don’t want to share.” P10, on the other hand, brought up their internal concerns around “losing the competition [with other companies] on building large language models,” by exposing too many model details to user auditors. As P10 shared at the end of the co-design session, “There are always things we can’t get and things we can’t share.”

5.5 Limitations

Similar to prior HCI work studying responsible AI practices in industry (e.g., [256, 343, 419]), our findings shed light on current practices, challenges, and needs among a set of practitioners who may be at the *forefront* of an emerging industry practice. As discussed in Section 5.3.1, we recruited participants using a purposive sampling approach [102]. All of our participants were passionate about addressing harmful behaviors in their AI systems, and they had direct experience experimenting with user-engaged approaches to algorithm auditing in their work. Furthermore, most of our participants worked at large technology companies, and all of our participants were located in the US (see Table 5.1).

5.6 Discussion

Drawing upon prior literature in areas such as algorithm auditing, crowdsourcing, participatory design, and fairness in AI, we discuss opportunities for future HCI research to help realize the potential and mitigate the risks of user-engaged auditing in industry practice.

5.6.1 Unique challenges in supporting user-engaged algorithm auditing

Prior research demonstrates great potential for human computation and crowdsourcing approaches to advance AI research and practice, especially in areas such as data generation and annotation [100, 438] and human-level evaluation [27, 538, 593, 594]. In our study, we found that industry practitioners’ current approaches to user engagement in algorithm auditing are often built atop existing crowdsourcing pipelines (Section 5.4.1). However, practitioners quickly ran up against limitations of conventional human computation and crowdsourcing approaches (Section 5.4.2). Below, we highlight five design implications for user-engaged auditing that extend beyond standard considerations for human computation approaches, and discuss corresponding directions for future HCI research.

Focusing on “who”

First, crowdsourcing approaches used to support AI research and practice typically focus more on *what crowd workers do* and less on *who they are* [27, 100, 438, 490, 538]. However, the “*who*” factors that encompass users’ intersectional identities and lived experience play critical roles in user-engaged auditing, as discussed in Sections 5.4.1 and 5.4.2. Prior work has also shown that users’ personal experiences with and exposures to bias influence the ways they search for and make sense of harmful behaviors

in algorithmic systems [91, 148, 165, 167]. Future research should explore better processes and tools to support practitioners in identifying and recruiting appropriately diverse and representative user auditors, for particular algorithm auditing tasks.

Supporting sustained, long-term contributions

Second, in contrast to the relatively transient, on-demand nature of most crowd work [65, 303], user-engaged auditing often requires sustained, long-term contributions from user auditors, to continuously improve AI systems (Section 5.4.2). This requirement entails radically different designs on both the process and interface levels. For example, future research may explore the interaction design space of in-situ feedback mechanisms, to solicit user auditors' feedback in the context of their day-to-day interactions with algorithmic systems. A key challenge for this line of exploration is to solicit feedback on algorithmic behavior in formats that are quick and relatively unobtrusive to collect from users, yet at the same time are readily interpretable and actionable for AI practitioners on the receiving end [43, 224]. To better support long-term contributions in user-engaged auditing, it is also critical to design intuitive and efficient interactive interface for users auditors [579]. In addition, future research may explore the design of social platforms to build sustained online communities of users who are motivated to engage in testing and auditing algorithmic systems together [474].

Navigating inherent ambiguities in auditing tasks and outcomes

Third, traditional crowdsourcing typically starts with well-defined goals and anticipated outcomes set up by the "requesters" [303]. In contrast, there may be benefits to empowering users to collectively take the lead in directing auditing efforts (Section 5.4.2). With too much direction and guidance from industry AI teams, user-engaged audits risk replicating the very biases and blindspots that they were meant to overcome (Section 5.4.2). Prior HCI research has mainly focused on developing tools and processes to better prompt users discovering more "unknown unknowns" [42, 505, 579]. Future research should explore the design of scaffolding mechanisms for user auditors that can navigate the trade-offs between promoting more effective algorithm auditing behaviors versus providing too much direction, limiting the kinds of issues user auditors are able to surface.

Reconsidering aggregation and quantification approaches to ensure marginalized voices are heard

In addition, as discussed in recent HCI research (e.g., [217]), typical crowdsourcing approaches involve aggregate analyses and evaluations, for example, by relying on "majority vote" from the crowd in order to arrive at the results. Practitioners emphasized that relying exclusively on quantification to derive actionable insights from user-engaged auditing could be harmful and counterproductive (Section 5.4.2). Resonating with findings from prior research [142, 342], we found that practitioners desired practical tools to support them in advocating for marginalized communities through an integration of both qualitative and quantitative forms of evidence (Section 5.4.3). Strategies like "tactical quantification" proposed by Irani et al. [271] could support practitioners in advocating on behalf of user auditors within their organizations, in industry contexts where numbers are culturally valued over more complicated stories. Future HCI research should explore the design of new tools,

computational techniques, and visualization approaches that can aid practitioners in persuasively advocating on behalf of marginalized groups of users (cf. [217, 432, 475]).

Designing user-engaged auditing mechanisms with teeth

Finally, prior work has emphasized the activist nature of auditing, which differs from much prior research on human computation and crowdsourcing [361]. The end goal for user-engaged auditing is to improve products and protect future users from algorithmic harms. Yet achieving this goal requires that companies are actually held accountable for addressing the issues that user auditors uncover [91, 244, 417, 458]. Thus, it is critical for future HCI research to design for user-engaged auditing with accountability and collective empowerment in mind. For example, researchers could explore building platforms that support user auditors in collectively applying pressure and holding companies to account, when serious issues are not addressed [271, 443]. In the next section, we further expand on the discussion around these design implications.

5.6.2 The complex relationship between industry practitioners and user auditors

Our study explored industry practitioners' perceptions of and relationships with user-engaged approaches to algorithm auditing. A key component of this is the relationship between practitioners and the user auditors who power the auditing process. As these two groups strive to surface and address harmful algorithmic behaviors, a complex relationship and a delicate *mutual (dis)trust* is revealed that can be friendly or antagonistic, depending on the situation.

How might users trust AI practitioners?

For user auditors, auditing is often seen as a form of activism, in which rooting out harmful behaviors in algorithmic systems benefits society [361]. Past research on user-engaged algorithm auditing highlighted users' advocacy for marginalized groups of people, expressing solidarity via their auditing activities [148]. However, industry practitioners in our study rarely brought up similar rationales for engagement with user-engaged auditing (Section 5.4.2). This might be because industry practitioners have greater opportunity to directly effect change on issues in the algorithmic systems they work on, whereas users typically need to rely on their collective power to raise awareness in order to be heard [361, 474]. Though user-engaged audits can empower users through the ability to directly connect with industry practitioners to try and identify issues together, they also firmly place the choice to take action with practitioners, potentially leaving users with less room for leverage via other means.

Given this asymmetric power dynamic between user auditors and AI practitioners, how could HCI researchers support empowering users' collective action when users' needs are not met and their trust fractured? Previous platforms like Turkopticon [271] and WeAreDynamo [443] demonstrated the potential for HCI researchers to consciously build spaces for activism, leveraging their collectiveness to negotiate their desires and needs. Incorporating similar spaces into future user-engaged auditing processes could alleviate users' concerns and empower them to act, ensuring that issues they collectively surface will be addressed in satisfactory ways.

How might AI practitioners trust users?

Despite the desire for holistic and critical auditing processes (Section 5.4.2), industry practitioners in our study expressed trust concerns around opening their systems to be audited by users (Section 5.4.3). As described above, users frequently turn to public awareness raising around problematic algorithmic behaviors they encounter as leverage to pressure companies into addressing those issues. How can practitioners fully trust users to audit their systems without publicizing issues before practitioners have had a chance to address them? Indeed, in our study, practitioners frequently cited apprehensions that user auditors might harm company reputations through negative PR (Section 5.4.3). However, other practitioners viewed themselves as more on the side of users. These practitioners work beyond their main job functions to advocate for users, while simultaneously using the issues brought by user auditors as evidence to convince their teams to act, putting pressure on companies from the inside. How might we support and protect practitioners who genuinely strive to mitigate or avoid problematic algorithmic behaviors in their AI products and services? To this end, similar to supporting user activism, there may be opportunities to build platforms (cf.[271, 443]) to support collective actions amongst these practitioner individual advocates within companies.

One possible avenue of exploration to address the tensions of required mutual trust might take inspiration from security bug bounties [347]. In these, security experts turn over information about security vulnerabilities to companies in exchange for monetary compensation and the promise of a fix, with the understanding between parties that if the issue is not addressed in a given time frame, then the security expert will publicize the issue. Borrowing from this model could enhance the trust between user auditors and industry practitioners as well, allowing practitioners a protected timeframe to fix issues but enacting a strict deadline for users to take further action if practitioners' promises are not upheld. Indeed, emergent projects like bias bounties [114, 418] have begun to transfer some bug bounty success to the territory of user-engaged algorithm auditing.

5.6.3 Users' perspectives on user-engaged algorithm auditing

As prior research has begun to explore users' practices and perspectives around user-engaged algorithm auditing [148, 315, 474], the main goal of our study is to begin to fill the gap between this emerging literature and industry practitioners' perspectives. However, while supporting practitioners in designing and implementing more effective forms of user-engagement in algorithm testing and auditing, it is critical to continue to explore users' perspectives and values. To this end, future work should bring together industry practitioners' and users' perspectives, potentially through the collaborative design, development, and oversight of user-engaged auditing procedures and platforms.

Importantly, when users are engaged in algorithm testing and auditing, they necessarily observe and likely experience some of those harms themselves. In our study, practitioners described thinking about better ways to target user auditors with relevant identity characteristics. This is in line with prior research, which has found that people with certain exposures and experiences are more able to surface related issues in algorithmic systems [148]. These people may be ideal candidates to serve as user auditors. At the same time, since these are often members of marginalized communities, who are already overburdened and more likely to be the targets of harmful algorithmic behavior [243, 265, 487], they are also more likely to be harmed

through participation in algorithm testing and auditing. Furthermore, drawing an analogy to platform content moderation, in which moderators are often exposed to violence and harassment and could be subject to long-term psychological harms [157, 499], auditing for problematic algorithmic behaviors may also result in long-term psychological harms towards user auditors.

Therefore, future research should consider and design to alleviate potential emotional burdens and psychological harms toward user auditors. Furthermore, user-engaged algorithm auditing could harm users if their labors are co-opted in ways that are not aligned with what they might have wanted. We highlight these burdens on users as vital areas for further research. Despite the burdens, user-engaged auditing, when implemented well, can serve to reduce algorithmic harms present and acting in the world now. We urge continual evaluation of user-engaged auditing processes by practitioners to ensure that these burdens and potential harms are mitigated. Future research should also explore the potential of computational or other alternative solutions to reduce the need for continuous involvement of the most vulnerable populations in the auditing process (e.g., [217, 315]), with the caution that computational approaches may also risk introducing new types of harms to users.

5.7 Conclusion

In this chapter, I conducted a series of interviews and iterative co-design activities with industry practitioners to gain insights into the current landscape and future opportunities for user-engaged algorithm auditing in industry practice. I surfaced major motivations for engaging users in AI testing and auditing and described practitioners' existing approaches for user-engaged auditing. I found that practitioners face challenges around appropriately recruiting and incentivizing user auditors, scaffolding user audits, and deriving actionable insights from user-engaged audit reports. Furthermore, practitioners shared broader organizational obstacles to user-engaged auditing, highlighting key tensions that arise in practice when involving users in algorithm auditing efforts such as potential PR risks, profit motives that work against protecting marginalized groups, and privacy and legal concerns. Based on these findings, I discussed the complex relationships between practitioners and user auditors, offering potential remediation for developing mutual trust. I then described various opportunities to support user-engaged auditing beyond existing design considerations for human computation or user feedback systems.

Bridge: Audits & Power

Chapter 6

Power-Related Pitfalls of Algorithm Audits: The need for challenging power structures

This chapter is adapted from my paper [150]:

Alicia DeVrio, David Gray Widder. “No More Plaudits for Algorithm Audits: A Provocation.” In *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*, (2026).

Algorithm audits have been increasingly proposed, pursued, and prioritized as a way to improve problematic algorithmic systems. And yet, the same issues continue to plague algorithmic systems, with algorithm audits making little headway toward remediation of algorithmic harm. In this provocation, we argue that algorithm audits are fundamentally insufficient mechanisms by which to ameliorate the violent impacts of algorithmic systems because they give the appearance of addressing algorithmic harms but do not meaningfully challenge the profitable systems they claim to scrutinize. To show this, we examine how algorithm audits sit within structures of power, bringing attention to three interrelated pitfalls of algorithm auditing: that algorithm audits too often *further* the concentration of power, that algorithm audits limit how both problems and solutions are envisioned, and that algorithm audits preserve a facade of positive change. We conclude with discussion of some necessary commitments for future work to start to grapple with and avoid these pitfalls and of how the insufficiencies of algorithm audits highlight broken dynamics in the broader AI accountability ecosystem. As such, we contend that fixing problematic algorithmic behaviors requires deprioritizing algorithm audits and instead shifting scrutiny to the systems of power in which audits and other accountability practices sit, which in turn mediate whether these practices and the theory of change they espouse are effective.

6.1 Introduction

Algorithmic systems have repeatedly displayed problematic behavior in ways that create, further, and amplify wrongful and structural harms as a result (e.g., [152, 362, 473]). Recent egregious cases include wrongful arrests following police use of facial recognition technology [341], algorithmic amplification of Nazi hate speech on TikTok [155], and algorithmic welfare payment discrimination against disabled people

and single mothers [358], among many others. In response to harms like these, algorithm audits have been proposed and pursued in attempts to remediate algorithmic system issues. Significant academic, industry, and government attention has focused on and prioritized algorithm auditing, leading the method to rise to prominence as a way to surface and counteract harmful algorithmic systems. For example, Safiya Umoja Noble's algorithm audit highlighted search algorithm discrimination, specifically that Google searches for "black girls" yielded overly sexualized results [47, 379, 380].

But are algorithm audits meaningfully fixing anything? Google has repeatedly claimed to address the issues in its search algorithm, but issues continue to plague the system. Noble herself describes how the initial searches that attuned her to issues within Google's search results "launched a years-long research process, following offenses and 'fixes'" (scare quotes original) [379]. In 2011, Noble identified the problematic search results for "black girls" in Google's search algorithm [379, 381]. While Google fixed this specific issue, in 2020 searching "Black girls" in Google's advertising platform resulted in algorithmically suggested adult content ads [39, 587]. And this came after Latanya Sweeney's 2012 audit showing how Google's advertising algorithm associated Black-sounding names with arrest record ads [512]—an issue Google had also already applied fixes to [214].

Cycles like this, from audit to issue to fix and back again, are not difficult to find. Google's translation service has gone in similar circles related to algorithmic issues of sexism (e.g., [75, 311, 407]). Auditing found racism in algorithmic predictive policing software PredPol in 2016 [489], but PredPol in 2019 was still used by many US cities [246]—and then in 2021 another audit found the same issues in the software [449]. In practice algorithm audits often surface and provide evidence for iterations of the same issues, issues that are already known, over and over.

In line with recent AI ethics scholarship interrogating the theories of change that guide the community's work [194, 388, 575], we are concerned about the theory of change upon which algorithm audits rely. The auditing landscape is going around in circles, repeatedly reinventing the wheel, inuring society to issues in algorithmic systems, both when they are brought up and when they are not sufficiently addressed. When society becomes accustomed to this circle—when the issues that audits find become normal and thus less shocking—audits lose their ability to spark change. Audits usually lack direct means of intervention [189], in turn degrading their utility as a tool to use to accomplish real, positive change.

Even when those in control have taken action to address algorithmic system issues raised by audits, the actions taken are often middling at best. Fixes are enacted in which little about the algorithmic system as a whole is changed, such as in the examples involving Google above, or those in power disregard the results of audits by arguing the audit's methodology is incorrect. And this latter response has come even in the case of very prominent audits. For example, after Virginia Eubanks detailed issues with the Allegheny Family Screening Tool (AFST) in the book *Automating Inequality* [170, 172], the Allegheny County Department of Human Services released a statement calling the audit inaccurate [171, 180]. Similarly, after ProPublica surfaced issues in Northpointe's COMPAS algorithm used to predict recidivism [32], Northpointe released a report that claims the audit's methods have errors [153]. Round and around the issues go.

At their core, algorithm audits are supposed to address issues in algorithmic systems—but what constitutes an algorithmic issue depends on perspective [571]. For those who interact with algorithmic systems or are otherwise impacted by algorithmic systems, the issues are the harmful, violent enactments of systems that were not

designed or built for them to thrive in [22]—that is, for impacted people, the issues are the negative impacts they are forced to experience: structural and wrongful harms. However, for those in control of algorithmic systems, such as tech companies, the issues are the complaints they receive about harmful algorithmic behaviors—complaints that bring the perception that their algorithmic systems perform poorly, the suggestion that they do not care about adverse impacts from their systems, and broadly bad PR. That is, for those in control of algorithmic systems, the issues are not the problematic algorithmic behaviors themselves but *the barriers those problematic behaviors create to their business incentives under capitalism*.

These differing perspectives of what constitutes an algorithmic system issue affect understandings of what algorithm audits are supposed to do, leading to a fundamental tension within algorithm audits. On their face, algorithm audits are designed and used to help surface, provide evidence for, and remediate algorithmic system issues. However, algorithm audits have been significantly prioritized by those already in control of algorithmic systems, and returning to their perspective of algorithmic system issues above, algorithm audits can then be understood as ways to reduce complaints about algorithmic systems and limit the resource investment they must spend on algorithmic system issues, all aligned with profit incentives. That is, algorithm audits serve to further the capitalist pursuits of those in control of algorithmic systems by convincing those affected by algorithmic systems that enough is being done to address the harms they encounter. Thus, audits are used to give the appearance of addressing harms, but do so in a way that does not meaningfully challenge the profitable systems they scrutinize. Or: algorithm audits support those with power in claiming to prioritize care while actually enacting prioritization of profit at the expense of those who use, interact with, and are affected by algorithmic systems.

Because of this, we argue that algorithm audits are insufficient, ineffective, and even harmful mechanisms to address the structural and wrongful harms of algorithmic systems. In this provocation we examine where algorithm audits sit within systems of power, bringing attention to three central pitfalls of algorithm audits manifested by this double bind: that algorithm audits further the concentration of power, that algorithm audits limit how both problems and solutions are envisioned, and that algorithm audits preserve a facade of positive change. We then outline some necessary commitments for future work as starting points for the AI ethics community to begin to confront and avoid these pitfalls of algorithm audits. These include suggestions that AI ethics community members orient themselves and their work toward tackling systems of harm, expanding notions of what is considered valuable evidence, acting directly and helping others act directly, and working to shift power to those affected by algorithmic systems. We conclude by discussing how the insufficiencies of algorithm audits highlight how the broader AI accountability ecosystem is broken, and call for the AI ethics community to thus deprioritize algorithm audits and instead shift scrutiny to the systems of power in which audits sit.

Our primary contributions are as follows:

- Three pitfalls developed from synthesizing and building on existing critiques that elevate these critiques into an urgent provocation that centers how broader power dynamics inhibit algorithm audits—and indeed the larger AI accountability ecosystem—from achieving lasting positive impact.
- A critique of and push to develop a better theory of change that drives how the AI ethics community handles problematic algorithmic systems impacts, along with some starting directions for how we as a community can begin to rethink

the existing AI accountability ecosystem such that its structuring better challenges power and holds those in control of algorithmic systems accountable for their systems' impacts.

6.2 What are algorithm audits?

In this work, we discuss algorithm audits and the theory of change they espouse. We align our work with Gansky and McDonald's argument that FAccT theories of change are ineffective, especially that they cannot be effective due to the dominance of profit and corporate power [194]. Following Birhane et al.'s critiques of algorithm audits as suffering from power asymmetries as well as limitations in scope, methodology, and practice [72] and following calls to focus more on power dynamics in AI ethics work [69, 71], our work questions whether addressing these issues is currently or ever possible under the banner of algorithm auditing due to audits' positioning within larger structures of power. In this section we explain our understanding of algorithm audits used in this piece.

Significant work has investigated and studied algorithm audits, which have been defined in a wide variety of ways [385, 541]. To name a few prominent examples: Sandvig et al. propose algorithm auditing as "a starting point for how researchers might look inside the black box of the algorithm to pursue knowledge about pressing public problems" [447]; Raji and Buolamwini describe algorithm auditing as involving "the collection and analysis of outcomes from a fixed algorithm" in order to "uncover problematic patterns in models of interest" [414]; Jack Bandy understands algorithm auditing as "an empirical study investigating a public algorithmic system for potential problematic behavior" [47]; Metaxa et al. describe algorithm auditing as "a method of repeatedly and systematically querying an algorithm with inputs and observing the corresponding outputs in order to draw inferences about its opaque inner workings" [361]; Birhane et al. and Ojewale et al. call for AI accountability to be a significant goal of algorithm audits [72, 385].

Guided by these dominant definitions in academic literature, we understand algorithm audits as systematic investigations of algorithmic systems to surface problematic issues, with the aim of achieving some form of meaningful accountability.

This begs the question: what is an algorithmic system? Algorithms and AI lack clear, universal definitions¹ [339, 503]; as such, we understand algorithmic systems based on their political implications. Ali Alkhatib has aimed to capture the broader project of AI by defining AI as "an ideological project to shift authority and autonomy away from individuals, towards centralized structures of power" [20]. We adopt this framing, gesturing toward the interplay between imaginaries and realities of algorithmic systems that can make powerful that which is seen as powerful [105].

Returning, then, to algorithm audits with this understanding of algorithmic systems in hand clarifies what the purpose of algorithm auditing. Broadly speaking, interventions against problematic algorithmic behavior that aim to be effective should ultimately counteract reductions of authority and autonomy, which we understand as problematic issues that, as per the above definitions, are currents embedded in the ideology of AI. As algorithm audits aim to surface and achieve meaningful accountability for algorithmic issues, algorithm audits as a project should ultimately halt or lessen the congealing of power inherent in algorithmic systems and shift power

¹Based on this, in addition to the frequent conflation of the terms and what they represent, we use "algorithms" and "AI" interchangeably throughout this work.

away from domineering institutions toward systemically disempowered individuals. In the following section we examine how algorithm audits are ineffective at accomplishing this.

6.3 Pitfalls of algorithm audits

In this section, we highlight three interrelated pitfalls of algorithm audits that surface upon examining how audits sit within larger structures of power. In doing this, we trouble conceptualizations of audits' abilities to address problematic, harmful issues in algorithmic systems. Our broad scope of inquiry examines algorithm audits within larger dynamics of power. The power-related pitfalls we draw out apply to external audits—arguably the only truly transparent and trustworthy forms of algorithm audits [125, 189]—as well as, in similar if not worse ways, internal audits. We note that distinctions between categories of internal and external algorithm audits can be complex and uncertain [519]. Despite work finding internal audits insufficient and calling for more external audits [334], in our examination external audits do not escape the core pitfalls we identify because they too are situated within and fail to significantly change larger dynamics of power in the AI accountability ecosystem (e.g., [204, 225, 578]).

This section provides insight into the following three pitfalls of algorithm audits: First, we describe how algorithm audits must be reckoned with as mechanisms that support the powerful in steering the narratives around negative impacts from algorithmic systems, in externalizing the labor required to address those impacts, and in retaining control throughout. Next, we highlight how algorithm audits frame problems too narrowly, constrain and exclude forms of knowledge, and are limited in their visions of solutions; that is, algorithm audits lack the expansiveness, on multiple fronts, required to sufficiently address issues arising from, perpetuated by, and intensified with algorithmic systems. Finally, we discuss how algorithm audits performatively reify and maintain the appearance of doing good, a facade that distracts from the insufficient capacity that audits have to actually drive change, which in turn leads to frequent reiteration of roughly known issues.

6.3.1 Algorithm audits further the concentration of power.

Algorithm audits serve to further solidify existing power structures, concentrating additional power in the hands of those who already have significant control over algorithmic systems [573], even as they purport to shift power to those with less. Using an algorithm audit is itself a choice that points in some directions and forecloses others, regardless of whether an algorithm audit can be characterized as more or less internal or external, and regardless of whether that choice is reducible to a single person, group, or institution. While internal audits inherently suffer from auditors coming from the bed of power, external audits are often beholden in other ways to power asymmetries and the interests of those in control of the systems they target [72], such as through nondisclosure agreements that restrict how much they can share about their findings [125, 385]. The prioritization of algorithm audits—any kinds of algorithm audits—thus furthers the concentration of power by instantiating a commitment to ways of knowing and ways of taking action that have been deemed acceptable and desirable within the larger currents of systemic power, often via people and institutions that hold considerable sway over those currents. We draw from DeVrio et al.'s framing of "algorithmlords" as "those holding greater

power to envision and construct algorithmic systems, occupying positions ‘above’ those subjugated by their decisions” [149]. For example, designers and developers could be *algorithmlords* while creating an algorithmic system, as could company management or city officials while deploying an algorithmic system, as could legislative members or other policy experts while requiring certain behaviors from an algorithmic system. We understand the designation of *algorithmlord* to apply in a way that varies depending on context and does not necessarily signal individuals’ endorsement of the systems they are part of, as even people in relatively powerful positions can lack the ability or knowledge to change these systems and their incentives [230, 572].

Algorithm audits orient additional power toward *algorithmlords* by allowing them to decide what is a problem in an algorithmic system. STS scholar Michel Callon has explored and described the power wielded in how an issue is “*problematized*” [101], and likewise, those who identify problems in algorithmic systems using audits have significant power to define the shape of the problem. Instead of allowing people affected by algorithmic harms to raise issues in ways that they choose, emphasizing parts of problematic issues they find most salient or urgent via their own means, algorithm audits place narrative control in the hands of *algorithmlords*: problem formulation occurs in specific ways due to auditing. To that end, issues are regarded as unknown or alleged until an algorithm audit documents them in a formal way, thereby relegating to the sidelines other efforts to identify and remedy issues that may occur without an audit.

Additionally, defining problems via algorithm audits empowers the definers to determine the solutions to those problems, as the shape of a problem lends itself more to certain directions forward. Narrative control of both algorithmic problems and their solutions by those already in power is further facilitated by power imbalances between relevant parties as well as by a society that has largely bought into engineering logic and systematic reasoning as more legitimate than other ways of knowing. Zong and Matias have pointed out in the realm of data collection that “it can appear to be a good thing when data collectors offer well-defined processes for people to opt out or voice dissent. However, data collectors often use these pathways of dissent to limit refusers’ agency by directing them away from other unspoken possibilities” [596]. Similarly, algorithm audits can appear to offer clear ways to raise harmful issues to the attention of *algorithmlords*, but they permit little agency for those affected and instead grant *algorithmlords* the power to direct when and how algorithmic systems are investigated, as well as in what ways issues are attended to.

More participatory forms of algorithm auditing can return some agency to affected people, but they contribute to the concentration of power in their own ways. While many algorithm audits do not even pretend to share or shift control of the process to those affected by algorithmic harm, forms of algorithm auditing have been proposed and pursued that include and empower communities and affected people in participation and some control of auditing processes (e.g., [141, 148, 289, 306, 474]). Algorithm audit processes that not only involve but are driven by on-the-ground communities have the potential to address some of the pitfalls we describe here, as evidenced by much work on public co-design of technologies and related technospaces (e.g., [259, 510]). However, algorithm audits that grant affected people and communities significant control to direct the process are far from the rule, and mere inclusion or participation in processes like algorithm audits should not be straightforwardly equated with shifts of power to those people [36, 70, 125, 487]. An especially pernicious aspect of participation like this in algorithm auditing, as described by the organizers of *Queer in AI*, is that it may “deny space for alternatives,

thereby serving as a justification of the systems in question” [16].

Algorithm auditing can also prop up existing structures of power by supporting algorithmlords in externalizing the work required to vet and improve their algorithmic systems. Participatory forms of auditing can exploit the labor of people in a system reliant upon trickle-down positive effects that neglects to ensure individuals are actually better off themselves as a result of their participation [70, 135]. Regardless of the level of involvement of affected people in an algorithm audit, the process often allows and supports relocating the burden of ensuring that algorithmic systems are safe, fit for purpose, ethical, fair, or otherwise good away from companies and others in control of these systems who should be responsible for their impacts and toward external groups who do the work of auditing.

6.3.2 Algorithm audits limit how both problems and solutions are envisioned.

That algorithm audits allow those in power to control the framing of both issues within algorithmic systems and potential ways to address those issues, as argued in the previous section, implies that algorithm audits severely limit the framing of both problems and solutions. While above we focus on how this consolidates power, in this section we discuss in what ways this shapes the problems and solutions themselves. Algorithm audits are mechanistically limited in their remit [72], a fact of design that confines their capacity to adequately identify and handle algorithmic system issues. That is, due to their formulation, algorithm audits are inherently small scope, individualizing, and reductionist, obscuring systems-level critique by restricting what is understood as acceptable evidence and focusing primarily on tractable problems and solutions. The mechanistic limitations of algorithm audits delegitimize types of issues that fall beyond the bounds of their inclusion as well as subjugate many potential futures—and their visionaries—through terms that make alternative possibilities incomprehensible. These pitfalls of overly repressed, deficient imagination for framing both problems in and solutions for algorithmic systems occurs across the spectrum of internal to external algorithm audits.

Algorithm audits constrain evidence by only making use of certain types of evidence. Algorithm audit evidence typically must allow for quantitative ways of viewing a potential problem, as many audits seek to statistically prove whether issues exist or are significant enough to warrant pursuing further. David Widder details how casting AI ethics work as objective, rational, and quantitative has legitimized it but has come at the cost of dismissing other kinds of knowledge that are challenging to fit within neat engineering-like processes such as audits [570]. Fundamentally this restricts what sorts of evidence is and is not acceptable for an algorithm audit. Qualitative forms of evidence such as personal anecdotes about experiences with an algorithmic system must be able to be categorized, measured, and compared. Similar to Sara Ahmed’s work describing how certain forms of complaints are legitimized by the infrastructures that exist for complaining [10], algorithm audits’ epistemic constraints on evidence further notions of what sorts of evidence is not only acceptable for audits but legitimate writ large.

In addition to constraining and problematically reshaping evidence, algorithm audits exclude other forms of evidence. For example, algorithm audits follow standardized forms of reasoning and logic, non-emotional forms including but not limited to seeking statistical significance as a vehicle for establishing proof. As such, they sideline emotional, embodied, and experiential forms of evidence that can also significantly highlight to what extent and in what ways issues can manifest in and

around algorithmic systems [298]. We relate this to the concept of respectability politics in which the presentation of issues and desires for change can be more successful when they adhere to notions of what is socially acceptable (e.g., [245, 251, 402]); similarly, algorithm audits advance specific notions of what is acceptable—that which is or can be systematically proven—and are designed to promote the surfacing and addressing of issues that work with and within this.

In part due to the limited nature of evidence they deem acceptable, algorithm audits frame algorithmic system problems narrowly. Methodologically algorithm audits focus on specific examples of algorithmic system failures that are straightforward to isolate, prove, and fix. This means that the audits are set up to show problems that have more tractable solutions, such as racial bias or disparate impact in an algorithmic system which can be addressed by balancing the discrepancy. In turn, this may contribute to why industry and others in charge of algorithmic systems have shown so much interest in audits, investing significant time and money into them in recent years. Indeed, audits provide natural solutions—but only ones constrained to the scope of the inquiry. By focusing on specific point examples of harm, audits are not a natural tool for documenting systemic problems emanating from algorithmic systems, of which there are many [22]. As Birhane et al. put it, reductions in problem and subsequently solution scopes can “risk equating processes and outcomes that are *fair* with those that are *equitable* and *just*” [71].

Overly narrow problem identification limits the reform made possible by audits to the logics of an already existing system. Algorithm audits are framed under the assumption that normally algorithmic systems are correct, innocent until proven guilty, and that issues revealed by audits are aberrations which can be individually removed. This makes the issues highlighted by people affected by algorithmic systems guilty until proven innocent: that is, the issues rather than the system bear the burden of proof and legitimacy. This move naturalizes algorithmic systems, forcing potential solutions to algorithmic issues to work within the assumption that the algorithmic system at hand will always exist, to take as predetermined the existence of the algorithmic system and its overarching logic. As such, algorithm audits cannot be broadly imaginative in their visions of alternative futures, of better worlds (see, e.g., [57, 236]).

6.3.3 Algorithm audits preserve a facade of positive change.

Finally, algorithm audits performatively purport to address problematic algorithmic behaviors but in reality frequently fail to accomplish this. This makes algorithm audits not only deceptive but even actively harmful themselves, as the prioritization of audits consumes valuable resources that could be used otherwise for positive change. Contributing to the appearance of positive change occurs regardless of whether algorithm audits are internal or external. While external algorithm audits avoid conflicts of interest and internal algorithm audits have potential to surface issues before deployment, both can provide reputation management for those in control of an algorithmic system [417]. That is, *algorithmlords* can gesture at any type of algorithm audit to signal that they are working to address issues from their systems, even when those audits fail to achieve lasting positive outcomes.

It follows then that algorithm auditing can be virtue signalling, when co-opted by powerful institutions for their own advantage. This can be considered a form of “ethics washing” whereby institutions like tech companies attempt to keep up the appearance of being ethical without actually substantiating with depth of action

[64]. This veneer of self-critique furthers the perception that positive change is occurring regarding harmful algorithmic systems while in reality enabling stasis [213]. Similarly, Lee Vinsel has used the term “criti-hype” to describe how some criticism of technologies can simultaneously generate interest in that technology and, more insidiously, distract from ongoing harms actively affecting people [554]. As Ahmed wrote about a complaint submitted to academic institutions: “In effect, the report was a means by which the university was able to not deal with the situation by appearing to have dealt with it” [10].

Beyond or perhaps because of this, algorithm audits at this point tend to reiterate issues generally known to exist. Racial and gender bias, for example, seem to repeatedly be “discovered” through audits of algorithmic systems. Seeking to prove harms rarely leads to adequate addressing of practices that cause those harms in practice [49, 71, 211], which can contribute to attempts to prove the same issues over and over again. Algorithm audits themselves often have little structural power to enforce changes based on their results; as such, their prioritization can result in focusing time, effort, and other valuable resource on surfacing or providing more evidence for known issues rather than working toward ways to address those issues, wider issues, and less commonly thought of issues.

In addition to resource seepage, algorithm audits can mask harms that come with the method. For example, some algorithm audits can lead to incremental, often quantitative improvements, such as the decrease in bias following the famous Gender Shades audit [414]; however, some biases still persisted years later, suggesting that larger changes than those offered by algorithm audits are required for justice [415]. Additionally, the algorithmic system being audited in that case involved gender classification, an oppressive technology itself (e.g., [235, 297, 416, 454]). This highlights how, beneath the surface of algorithm audit fixes, there often lie more fundamental human rights issues with the algorithmic system that algorithm audits are unfit to handle (e.g., [247, 299]). To that end, Keyes and Austin have described how auditing can produce new harms even while aiming to address existing harms [298]. In garnering evidence of harm, algorithm audits often ask survivors of trauma from algorithmic systems to relive their experiences and translate them to be legible for others who may be more invested in the audit than in advocating for those affected or otherwise working toward justice or harm reduction.

6.4 Ways Forward: Grappling with power toward a more effective theory of change

In this provocation, we have argued that algorithm audits are insufficient, ineffective, even detrimental ways to address the structural and wrongful harms of algorithmic systems. To do this, above we have examined some significant, power-related pitfalls of algorithm audits. In this section, we provide directions forward for the AI ethics community to reconsider the algorithm auditing theory of change and reorient future work in light of this, all via attending to and grappling with power dynamics. We start by providing a set of commitments the AI ethics community must take on in order to begin to avoid the pitfalls we identify above. Then, we discuss how the pitfalls of algorithm audits highlight broken dynamics in the larger algorithmic accountability ecosystem, and consequently we encourage restructuring this ecosystem such that more radical futures are prioritized and those with power over algorithmic systems are held accountable for their systems’ impacts

6.4.1 Embracing new commitments

In this section, we suggest necessary commitments for the AI ethics community to take on such that they can begin to grapple with and avoid the above pitfalls. These commitments are not intended to be comprehensive fixes and do not perfectly correspond to the pitfalls described above; rather, we see these as fundamental orientations that, when truly pledged and pursued, might help in future avoidance of algorithm audit pitfalls. We seek to encourage pragmatic directions that on the one hand do not accept sociotechnical systems as they are, as unable to be changed or controlled [406], while on the other do not ignore that existing structural currents shape the potential impacts of any actions we undertake.

First, we entreat the AI ethics community to commit to *tackling systems of harm* rather than point instances of problematic algorithmic behavior. Algorithm audits tend to focus on individual issues within algorithmic systems; as such, we see opportunities for expanding focus to encompass entire algorithmic systems, including the factors that contribute to their design, development, and deployment in the world as well as their internal working. This has been understood as a commitment to “non-reformist reforms” which, in contrast to reformist reforms that look for ways to improve existing systems, look for “ways to achieve emancipatory social conditions” by starting beyond rather than within existing systems [219, 221, 266]. The incremental changes that come from reformist reforms might be seen as positive on their face but in reality can reinforce oppression long-term by supporting the status quo [221]. As such, we echo other scholars in calling for algorithmic issues to be engaged with as sociotechnical problems that can only have sociotechnical solutions (e.g., [182, 465, 473]). This acknowledges how algorithmic systems can be imbued with harms as a feature [236], requiring that their entire logics and existence be brought under question for adequate mitigation of their issues.

Second, we call for the AI ethics community to commit to *expanding notions of what is considered valuable evidence* beyond the systematic, quantitative nature of evidence that is legitimized by algorithm audits. Scholars have emphasized the importance of translation work [374, 508]; we similarly see opportunities for researchers to convey evidence of algorithmic system issues from affected people to those with the power to address those issues in ways that are comprehensible and understood as significant to those in power. We understand that translation work can be conducted via algorithm auditing where necessary. In many cases, however, we see opportunities for the AI ethics community to work to convey the legitimacy and importance of varied forms of evidence to those in power regardless of what the existing norms of legitimacy and importance might dictate. We especially see the need for advocacy for human aspects of algorithmic impacts—often excluded by algorithm audits—as evidence. For example, emotional evidence conveying how algorithmic issues made people feel or acts of anger in response to problematic algorithmic behavior can be understood as legitimate and productive feedback about in what ways and to what extent algorithmic systems have issues that can be addressed. Including and elevating these sorts of evidence can expand legitimacy to more of the ways that everyday people engage with and are affected by algorithmic systems as well as support the redirection of resources toward addressing issues rather than toward seeking to find additional proof.

To that end, we encourage AI ethics community members to commit to *acting directly and helping others act directly* rather than working to convince others to act to improve conditions created and exacerbated by harmful algorithmic systems. The anarchist principle of direct action provides a way to make change without waiting

for the approval or buy-in of powerful institutions or people and has been proposed as useful for the realm of AI ethics [7]. Affected people already frequently respond to algorithmic harm they encounter using what power they have in a variety of ways [149, 313, 568]. For example, destruction and sabotage of algorithmic systems has been observed and even promoted as a way to more comprehensively address harmful algorithmic systems [21, 149, 313, 552]. Relatedly, we highlight that anger can be productive and a form of power [11, 46, 113, 336]. Though algorithm audits relegate emotions to the sidelines of what justifies evidence—as discussed above—as well as what is useful for building power, we see anger as having the potential to bring groups together for collective action, to motivate people to have their voices heard, and to encourage those in charge of algorithmic systems to listen and make tangible change in response. For example, anger can spur significant negative PR, which can be a strong incentive for tech companies to address issues (e.g., [149, 474]). Relational ethics also offers pathways toward more just outcomes in and around algorithmic systems [69]. For the AI ethics community, this can mean prioritizing work and research that supports people affected by algorithmic systems in what they are already feeling and doing to resist harmful impacts of those systems [267]. This could involve creating networks to connect people for collective action and solidarity, designing pathways for raising issues in more visible ways, supporting the development of more localized sociotechnical systems that do not rely on big tech, or building tools to allow people to more fully opt out of harmful systems or avoid parts of harmful systems they do not want to engage with.

In this vein, we call on the AI ethics community to commit to *shifting power to those affected by algorithmic systems* in ways that cede more control than many forms of participation do [9, 243, 284]. Supporting everyday people in taking action directly, as above, is a productive way to do this. Auditing has origins as a community-driven method for accomplishing real change [541]. More priority from the AI ethics community can and should be given to what people affected by algorithmic systems are themselves prioritizing [510]. The AI ethics community can reprioritize, shifting focus from finding and proving issues via audits to understanding and supporting what people are already doing to resist and shape algorithmic systems to create better on-the-ground outcomes [506]. This means supporting community ownership and control of algorithm audits instead of bringing in community members as merely included consultants, participants, or people who labor to surface issues. Communities need to have the power to redirect and reprioritize algorithm audits, which necessarily requires AI ethics community researchers and others to cede some of their own power to make those decisions in return. This also means allowing community members to define the frame of what should be done with regards to algorithmic systems, such that the possibility space can include the refusal to engage with, the decision to destroy, or the creation of joyful alternatives to harmful algorithmic systems as well as survival in and mitigation of harmful algorithmic systems. We encourage the AI ethics community to look toward the many community collectives that already exist and to offer themselves as partners, collaborators, and assistants to accomplish their aims.

6.4.2 Scrutinizing the accountability ecosystem

The existing structures of accountability surrounding algorithmic systems, especially those related to the regulation of tech companies, often necessitates the forms of language and evidence provided by algorithm audits in order to convince those in power to begin to make changes. Though we argue that algorithm audits have

pitfalls that make them ineffective at addressing problematic algorithmic behavior, we would be remiss not to acknowledge that they can be valuable tools in some ways. Audre Lord famously wrote, “For the master’s tools will never dismantle the master’s house. They may allow us to temporarily beat him at his own game, but they will never enable us to bring about genuine change” [335]. Following David Widder’s work on hierarchies of epistemic power in AI ethics labor [570], we identify the master’s tools as not the audits themselves but as the ways in which audits instantiate hegemonic perspectives about worth: what is worth listening to, what is worth taking action on, who is worth supporting. When audits are leveraged to counter these perspectives, such as through using audit reports to pile on evidence to showcase to those in charge of algorithmic systems, they can eventually lead to incremental benefits.

But there is a sense of stasis in finding and proving very similar issues repeatedly, a sense of helplessness in seeing this as the primary vector for changing an issue-strewn technological landscape. To free us from what can feel like and in many regards is a Sisyphean approach, we look toward ways to expand the notion of what algorithm audits are. Existing research has sought to critique algorithm audits and broaden the scope of what we can consider to be an audit beyond more expert- and/or institutionally-led investigations that focus on systematicness and scientific proof above all (e.g., [99, 201, 306, 474]). One piece at a time, we can tactically replace the parts of algorithm audits while remaining under the banner of algorithm auditing, trading sanctioned representations of harm with more expansive ones that better match affected people’s experiences, until we have created an audit of Theseus that shifts the onus of scrutiny from affected people providing evidence for the harms they experience onto those with control over the algorithmic systems that instigate those harms, as it should be.

Ultimately, we view algorithm audits as “tactics of soft resistance” [377, 576], holding some possibilities but reliant upon the discourses and logics of the systems they try to make change from within. But what power might we gain from freeing our imaginations from dominant ways of thinking [57], beyond “soft” audits? We advocate for the abolition of auditing—or the expansion and repositioning of the notion of auditing within the larger ecosystem of accountability, such that audits become “hard”—in favor of more radical tactics that shift power more wholesale to those who are affected by algorithmic systems’ negative impacts, allowing them to set the terms of what is required for their own survival in the landscape of algorithmic violence and rejecting the notions of legitimacy that current structures of power insist upon. Building on calls to strengthen AI accountability infrastructure (e.g., [41, 125, 385, 417, 418]), we contend that we should direct much of our attention as a community of research and practice toward finding ways to reshape this ecosystem’s power dynamics and finding, reckoning with, and moving in directions that prioritize futures with more imaginative possibility that are unconstrained by or actively refuse the logics and requirements of algorithmlords today. This will be a significant undertaking that requires collective input and effort, interrogation and reworking of stagnant theories of change, and embracing and taking action with the new commitments above in hand to better challenge power.

Locating responsibility with the powerful. The starting commitments we point toward and indeed our argument critiquing the effectiveness of algorithm audits largely focus on the AI ethics community and those affected by algorithmic systems. We have called upon the AI ethics community to reorient its stance toward algorithm audits due to fundamental flaws in the theory of change they pursue. In doing so, we advocate for increased engagement with and power ceded to communities affected

by algorithmic systems. These directions are given assuming that those in control of algorithmic systems cannot be relied upon to adequately address the harmful impacts of their own systems.

However, by focusing on affected people and their determinations of what are reasonable ways to remedy algorithmic issues, we have implicitly but unintentionally put some burden of responsibility on them to determine those pathways forward. For instance, expanding the scope of auditing likely still requires affected people to identify issues they encounter, which can burden and even involve harm to those people in the process [532]. A tension exists here, as affected people attempt to drive change because the status quo is that their experiences and lives are worsened by harmful algorithmic systems. Only when people affected by algorithmic systems also have power and control over those systems themselves can this tension start to be resolved—that is, when they are in positions such that they engage not only because their survival necessitates it but also to create technospaces of radical joy and thriving for themselves.

To reach this future, we emphasize that responsibility for algorithmic system impacts, and especially for algorithmic system issues, must remain with those with power over algorithmic systems: the organizations and institutions with control of the design, development, and deployment of these systems. We echo Lucy Suchman’s framing of “located accountabilities”, or how those involved in the production of technology should identify, understand, and take responsibility for their positions within larger structures of power [502]. At the end of the day, while we look to the potential of systemic change starting now, those with power over algorithmic systems should remain accountable for the outcomes of those systems. Let “soft” audits not placate and distract us from organizing for refusal, demanding hard regulation, and raising hell until change occurs.

6.5 Conclusion

In this provocation, we have examined three power-related pitfalls of algorithm auditing that illustrate the ineffectiveness of algorithm audits’ espoused theory of change. Looking forward, we call for a reorientation of the AI ethics community that involves commitments to grappling with power dynamics and reprioritizing such that algorithm audits as they are now are no longer pursued as a primary way to address harmful algorithmic systems. We emphasize that the AI ethics community should rethink the structure of the AI accountability ecosystem and direct more focus, attention, and material resource toward supporting and themselves participating in the work that affected people are already doing to address algorithmic system issues.

Part II

Power Dynamics around Bottom-Up Resistance

Chapter 7

A Taxonomy of Responses From Below to Algorithmic Harm

This chapter is adapted from my paper [149]:

Alicia DeVrio, Motahhare Eslami, Kenneth Holstein. “Building, Shifting, & Employing Power: A Taxonomy of Responses From Below to Algorithmic Harm”. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. (2024).

A large body of research has attempted to ensure that algorithmic systems adhere to notions of fairness and transparency. Increasingly, researchers have highlighted that mitigating algorithmic harms requires explicitly taking power structures into account. Those with power over algorithmic systems often fail to sufficiently address algorithmic harms and rarely consult those directly harmed by algorithmic systems. Left to their own devices, people respond to algorithmic harms they encounter in a wide variety of ways, but we lack broader, overarching understandings of the various methods people engage to resist harmful AI systems. In this work, I synthesize documented, historical cases into a taxonomy of responses “from below” to algorithmic harm. This taxonomy connects different types of responses to existing theorizations of power from fields including anthropology, human-computer interaction, and communication, centering how people employ, shift, and build power in their responses to algorithmic harm. Based on the taxonomy, I highlight an opportunity space for the FAccT community to engage with and support such action from below.

7.1 Introduction

Aiming to root out algorithmic harms, a large body of research has attempted to ensure that algorithmic systems adhere to notions of fairness and transparency. Increasingly though, researchers have highlighted that effectively addressing harmful algorithmic behaviors requires centering systems of power in and around algorithmic systems [22, 69, 74, 284, 299, 364]. This focus is especially important since existing structures around algorithmic systems often fail to empower those who directly interact with and are affected by algorithmic harm [22].

In response to such power imbalance, people affected by algorithmic harm frequently adopt workarounds and counter-practices to algorithmic systems in place. Recent research and real-world case studies indicate that responses “from below”—i.e., from members of the public or others without direct power to make decisions about

the design or deployment of algorithmic systems [568]—can be effective in garnering power and working toward harm remediation [148, 313, 474, 552, 568]. For example, Twitter users who noticed, tweeted about, and extensively worked together to test racial bias in the platform’s image cropping algorithm succeeded in causing Twitter to reconsider and change its use of the algorithm [474]. And after Uber and Lyft algorithmically determined inadequate ride fares, drivers coordinated turning their apps off to feign a driver shortage and were able to boost fares [313, 513].

While responses from below can be effective in mitigating some algorithmic harms, actors’ structural disempowerment can limit the effectiveness of their responses. For example, Elon Musk disbanded Twitter’s AI ethics team after taking over the company [140], and rideshare companies continue to perpetuate algorithmic wage discrimination [160]. How can we better understand the ways that responses from below interact with power? What lessons can we draw from past responses to empower future action from below?

In this chapter, I develop a taxonomy of existing types of responses to algorithmic harms, centering actions that people take from below in response to algorithmic harms. I focus on the nature of the responses themselves and, where described by actors, their motivations and aims; I do not in this work attempt to categorize the effectiveness of various responses. Our taxonomy emphasizes how these responses from below are linked to power dynamics through three broad categories: “*protecting*” themselves by employing their own power, typically within the context of a harmful algorithmic system; “*pressuring*” those in control of algorithmic systems by shifting power away from them; and “*strengthening*” their capacity by building their reserves of power, enhancing their abilities to pressure and protect. Protecting responses can lead to mitigations more immediately, but they tend to occur within the existing logics of algorithmic systems, which inherently limits their outcomes. For example, when users found Booking.com’s hotel review algorithm problematic and acted to fix its outputs by manipulating their review scores, they were constrained by fundamental issues within the algorithm: its lowest possible output rating was 2.5, so altering inputs did not help when users wanted ratings lower than 2.5 [168]. More widespread, systemic change often comes from pressuring responses. For example, following worldwide protests of unfair algorithmically generated A-level exam scores, the algorithmic scores were revoked and replaced with teacher-determined ones [404]. And people need sufficient power, which strengthening responses build, to do any of this. For example, to escape algorithmic surveillance and control, food delivery workers in China created virtual communities that allowed them to share tactics to better navigate food delivery platforms [509].

Based on this taxonomy of responses, I discuss opportunities for the FAccT community to better engage with everyday peoples’ current actions in response to algorithmic harm to further amplify actions they are already taking in productive ways. This work identifies ways that structurally disempowered people create and further their own counterpractices intended to secure power over algorithmic systems. Supporting efforts from below can put those most affected and with little systemic power in control of remediation efforts, which can enable progress even when powerful actors like companies fail to see worth or incentivization in working toward algorithmic harm remediation. In addition to efforts to mitigate harms from above (e.g., by supporting developers in assessing and addressing issues), it is critical that the FAccT community helps affected people defend themselves from actively harmful systems.

7.2 Related Work

7.2.1 Responses From Below to Algorithmic Harm

Significant research investigates how users and others affected by algorithmic systems perceive algorithmic harm (e.g., [148, 192, 328]). For example, DeVos et al. sought to understand how users identify, make sense of, and evaluate potentially harmful algorithmic behaviors [148]. In this work, I focus on steps that everyday people take after they have perceived harm in an algorithmic system.

Recent research has revealed that people with little structural power have unique abilities to respond to algorithmic harms [148, 313, 474, 552, 568]. With the prevalence of algorithmic harm, many of these people's responses can be considered forms of "everyday resistance", a continuous struggle implemented with "the ordinary weapons of relatively powerless groups" [460]. I also draw on "routine infrastructuring", as actors work to re-appropriate and build resilience around algorithmic systems [466]. For example, the concept of "everyday algorithm auditing" highlights how users, in their everyday interactions with algorithmic systems, can surface, interrogate, and *work toward remediation* of harmful algorithmic behaviors that more formal auditing methods fail to find [474].

However, we lack broader, overarching understandings of the nature of the existing responses from below to algorithmic harm. Some work has investigated the ways that users and others affected respond to harmful algorithmic behaviors (e.g., [168, 313, 485, 596]). Throughout, many questions have arisen regarding the nature of these remediative responses from below. In this chapter, I start to provide this insight, extending past literature around responses from below such as data leverage and POTs [313, 552]. However, instead of centering data or technologies, this work centers what the disempowered do now to respond to algorithmic harms and how their actions are rooted in dynamics of power.

7.3 Taxonomy of Responses from Below to Algorithmic Harm

In this section, I describe nine types of responses from below to algorithmic harm, organized into five broad categories (see Table 7.1). Our taxonomy centers *people* and captures the actions they take from below to *employ, shift, or build power* in response to algorithmic harms. Each subsection presents a *type of response* and examines power-related *strategies* associated with it, using a range of existing examples to illustrate the scope of each.

7.3.1 Process of Response Collection & Categorization

Our research team gathered 96 documented cases with 169 embedded responses from below that intend to remediate perceived algorithmic harm to create our taxonomy. The collection process started with an initial set, and expanded via iterative search of various resources including research literature, news media, and social media. Our cases spanned many different algorithmic domains (e.g., recommendation, image cropping, facial recognition) and types of harm (e.g., financial, stereotyping, surveillance). More collection details can be found in Appendix B.

I understand actors and actions from below as those lacking direct power to control the design and deployment of algorithmic systems. To emphasize power relations, I broadly refer to those who have control over algorithmic systems as "*algorithmlords*", inspired by past work describing "data barons" who control data and

Response Type ↓	Protecting	Pressuring	Strengthening	Examples
Mitigating individual algorithmic harms	X			Changing search terms after problematic results [148], Engaging with hidden content to amplify it [288]
Pursuing legal avenues against algorithmic harm		X		Instituting policies to reclaim AI governance [183, 569], Litigating companies who deploy harmful algorithms [169]
Investigating potentially harmful algorithmic behaviors further			X	Researching background information [568], Testing potential issues [474]
Refusing engagement with harmful algorithmic systems				
Impairing an algorithmic system	X	X		Slashing tires of self-driving cars [421, 434], Unplugging privacy-invasive cameras [228, 592]
Depriving an algorithmic system	X	X		Logging out of Facebook [561], Leaving data-labelling job [398]
Confusing an algorithmic system	X	X		Mirroring, cropping, and filtering duplicate videos [584], Downloading software that automatically clicks ads [264]
Communicating algorithmic harm with others				
“Complaining” to those with more power	X	X		Contesting errors using a platform’s appeals system [302], Submitting complaint to government agency [104, 120]
Publicizing algorithmic harms broadly		X	X	Tweeting about an issue [176, 474, 529, 590], Holding a press conference [568]
Sharing, discussing, and organizing with others affected by the algorithmic system	X		X	Describing in writing where others can see [168, 474], Creating, joining, and discussing in WeChat groups [591]

TABLE 7.1: Types of responses (bold), their strategies that interact with power (X), and examples

“digital feudalism” in which tech companies behave as modern-day feudal lords [159, 205, 275]. In the vein of [568], I understand *algorithmlords* as those holding greater power to envision and construct algorithmic systems, occupying positions “above” those subjugated by their decisions, while acknowledging that relationships between people are more complex and nuanced than this fully captures (see [164, 253]).

I extracted and qualitatively coded responses from the cases I gathered, then conducted a bottom-up thematic analysis with a series of interpretation sessions on these responses. Based on this, I grouped the responses into *types* of responses based on similar responder behaviors present and, based on prior literature on power, synthesized these into three high-level *strategies*, each of which captures a different way that these responses aim to interact with power:

Protecting strategies: ways that people responding to algorithmic harm act from below to directly affect their or others’ “power-to” do [456]. Typically this occurs within the context of the algorithmic platform itself and serves to make people’s interactions with algorithmic systems safer. These responses fit into repair politics [262, 273]—that is, responses that work “through improvisations, patches and ingenuity, together with and within algorithmic systems” [542]—and more immediately improve the experiences and abilities of the disempowered.

Pressuring strategies: ways that people responding to algorithmic harm act from below in attempts to gain “power-over” *algorithmlords* [456]. These responses often have the ultimate aim of persuading *algorithmlords* to make changes. However, they may vary in their degree of forcefulness, from punishing an organization with negative PR, to convincing an organization by flagging issues for them to address,

to forcing a company with new legal requirements.

Strengthening strategies: ways that people build up their reserves of power, enhancing their future abilities to protect or pressure. Strengthening also encompasses ways that actors recognize and become better equipped to leverage their existing, often collective, power. Strengthening can help people understand the scope of the harm they have encountered, determine what should be done next, and gain perceived legitimacy.

I see the cases used as the basis for our taxonomy as illustrative but not necessarily exhaustive. Responses and types of responses commonly span multiple strategies. Additionally, due to the constantly changing and emerging nature of algorithmic systems, of their applications, of related harms, and of related responses to these harms, I see this work as an initial exploration and invite future researchers to expand on it. More process details can be found in Appendix B.

7.3.2 Mitigating Individual Algorithmic Harms

In this type of response, people adjust their own behaviors in and around algorithmic systems to change how the system works in a particular circumstance, without fundamentally changing how the algorithmic system works as a whole. I understand such actions as *protecting* responses, as people employ their own abilities to shield themselves from the harms a given algorithmic system foists upon them. These responses parallel conceptions of technological repair, which has been theorized as “the ongoing work of fixing and maintaining the objects and systems around us” [262] and has been previously applied to users addressing harmful algorithmic behaviors [168]. Additionally, these responses can be seen as a form of algorithmic resistance, theorized by Velkova and Kaun as action that “takes place from *within* the logic of the algorithm” and “does not deny the power of algorithms but operates within their framework” [542].

Algorithmic harm workarounds. Sometimes people act to mitigate individual algorithmic harms by avoiding them, creating distance between themselves and potential harm (e.g., [420, 486]). For example, Chinese food delivery workers chose not to follow platform-suggested routes, instead relying on their own knowledge of faster shortcuts [591]. In another instance, Bangladeshi users who felt misunderstood and unreasonably censored by Western Facebook moderation policies changed the ways that they posted—adding English translations, combining Bengali and English letters to write, and refraining from posting publicly [467].

Avoidance can also occur before harmful algorithmic behavior presents. A Google Images user anticipated a heterosexual bias in “wedding” search results and preemptively switched to use the search term “lesbian wedding” from the start [148]. As another example, Amazon began using AI Netradyn cameras that frequently punished delivery drivers for unsafe driving in situations that were not unsafe or not the driver’s fault. Some drivers responded by driving over cautiously to preemptively avoid punishment, such as braking “once before a stop sign for the Netradyn camera, and another time for visibility before crossing an intersection” [229].

Instead of changing their usage to operate around harms, some actors protect themselves by masking algorithmic harm from their view. This can present in the form of using built-in platform affordances to block certain content or people [148]. For example, TikTok users pressed “not interested” buttons associated with content they did not wish to see or found violating [485], and Bangladeshi Facebook users who saw problematic advertisements responded by unfollowing related pages as well as muting ads using the platform’s settings [471]. Other times, this masking is

less supported by platforms and their affordances. Ad blocking is a prevalent example of this, wherein users who dislike the experiences they receive from advertising algorithms use third-party blockers to remove the issues from their experiences [354]. TikTok users also sometimes protect themselves by intentionally seeking out and engaging more with videos similar to what they want to see, in order to shape algorithmic behavior [485].

Beyond self-protection. The workarounds above tend to protect only the person responding. In the above examples, others can be impacted secondarily, but these effects can be difficult to identify due to their diffuse nature. For example, if enough TikTok or Facebook users mark a specific post as uninteresting, the recommendation algorithm might deprioritize that video preemptively on other users' pages.

People also directly act to mitigate individual algorithmic harms that can protect many others beyond those taking action. For example, TikTok users who believed that the For You Page algorithm suppressed videos with social-justice content responded by engaging with apparently suppressed videos en masse: users commented, liked, and shared these videos repeatedly in attempts to boost video content that might have otherwise been hidden [288]. These actions aimed to amplify content broadly on the platform, using the For You Page algorithm to spread potentially previously suppressed content to many other TikTok users. And social workers using an AI tool individually adjusted and compensated for biased and incorrect suggested child-maltreatment risk scores, with the goal of protecting children and families [112, 293]. Similarly, Indian community health workers re-ran tests when they believed an AI app's diagnosis was incorrect [386].

Other times, actions to mitigate algorithmic harms may protect a subset of users. For instance, in response to beliefs that the For You Page algorithm suppressed videos created by people of certain identities, such as LGBTQ+ and Black creators, some TikTok content creators with more algorithmically favored identities shared their accounts with suppressed creators, allowing amplification of algorithmically suppressed voices via new platforms [288]. Additionally, some TikTok creators whose videos were suppressed responded by simply re-uploading videos that were taken down or by changing the aesthetics of their videos to better fit what they believed the algorithm prioritized [288, 485]. These actions aimed to amplify content mainly to the set of TikTok users who engaged with a certain content creator, either an algorithmically favored one in the former example or an algorithmically suppressed one in the latter ones.

7.3.3 Pursuing Legal Avenues Against Algorithmic Harm

In many cases, actors from below take legal action in response to encountering algorithmic harm. In this type of response, people attempt to leverage legal institutions to shift power away from algorithmlords; thus, I consider these responses to be a form of *pressuring*.

One way people do this is by suggesting and signing bill initiatives that aim to curb harmful algorithmic behaviors that they have encountered. For instance, three Californians, upset with how technology companies collect and use their personal data in algorithmic systems, proposed a ballot initiative for consumer privacy [556]. More than 600,000 other Californians signed in support of the initiative, certifying it so that it could later pass into law [555, 556].

People also exert pressure via direct legal action against algorithmlords. As one example, many business owners, believing that Yelp algorithmically removed positive reviews from their profiles because they did not pay for Yelp's advertising, filed

almost 700 Federal Trade Commission reports against Yelp [169, 474]. Additionally, Amazon has recently been sued for “selling suicide kits”, or algorithmically recommending buying combinations of items expedient for poisoning oneself, by parents of users who have died by this method [33, 250, 529].

Finally, due to the concentration and predominance of *algorithmlords* in colonialist (often Global North) countries, entire jurisdictions (often at some point colonized and/or Global South) occupy positions of subordinate power in and around algorithmic systems. As such, governments of these nations can respond from below to colonialist, exclusionary algorithmic systems by creating new policies to force *algorithmlords* to behave in ways less harmful to them. Indigenous Data Sovereignty is one example of this, advocating for “the right of Indigenous peoples to control data from and about their communities and lands, articulating both individual and collective rights to data access and to privacy” [403]. Indigenous groups worldwide have adopted processes and principles supporting this, such as the Māori in Aotearoa/New Zealand, the Sámi in Sweden, and the Ktunaxa Nation in Canada [413]. In another example case, much African population data must be stored in servers outside the continent since non-African companies own much of the data infrastructure; in response, African governmental leaders are instituting policies to make these companies adhere to privacy regulations and pay taxes, shifting some of the governance to themselves [163, 183, 403, 543, 569].

7.3.4 Investigating Potentially Harmful Algorithmic Behaviors Further

Another type of response from below to harmful algorithmic behaviors involves gathering more information about an issue. For instance, when Rotterdam’s welfare fraud algorithm began unreasonably ranking people as high risk, one affected person requested the information contributing to their score [94]. This information can be leveraged to add legitimacy to people’s claims and convince other groups with more power to listen and take action as well. As these investigations primarily help disempowered people build power, I view this response as *strengthening*.

Investigation can build people’s power by helping them figure out what to do next and legitimize their claims. For example, San Diego residents dedicated significant effort to reading documents related to proposed smart streetlights as a “fact-finding mission” [568]. These people then leveraged what they had learned to refine their arguments and ensure use of language legible to those they engaged with [568].

Shen, DeVos et al. introduced the concept of “everyday algorithm auditing” to describe the process by which people “detect, understand, and/or interrogate problematic machine behaviors via their day-to-day interactions with algorithmic systems” [474]. For example, users conducting an everyday algorithm audit of potential racial bias in Twitter’s image cropping algorithm carried out tests and gathered evidence around various hypotheses that they developed [474]. This helped users assess the extent to which racial biases were present and build up evidence that they used to push Twitter for change. As reviewed in [474], numerous cases of everyday algorithm audits have been documented in recent years.

7.3.5 Refusing Engagement With Harmful Algorithmic Systems

In another type of response to algorithmic harm, people refuse to engage with harmful algorithmic systems. These responses draw on concepts of Indigenous and feminist refusal [12, 483] and especially parallel refusal of various technological practices [118, 195, 595, 596]. Thus, I consider these responses to be forms of *algorithmic refusal*.

Often people do this by removing some or all of themselves from an algorithmic system's visibility. This type of response can be *protective* in the sense that it requires people to employ their own power and, typically, limiting or eliminating true engagement serves to expose people to fewer harms, which makes their interactions safer. Additionally, this type of response can *pressure* by diminishing the engagement and data that algorithmic systems can consume. This penalizes those in charge of algorithmic systems and shifts some power and control to people acting from below.

People can also refuse engagement by actively seeking to impair or confuse algorithmic systems. These types of responses serve to *pressure* algorithmlords, making their disapproval of a system hard to ignore, and to *protect* themselves, obstructing an algorithmic system from its typical functioning.

Impairing an algorithmic system

Disempowered people sometimes react to harmful algorithmic systems by attempting to render them useless. For example, upset at being part of Waymo's self-driving car beta tests without their consent and at injuries these cars had caused, Arizonan pedestrians chose to fight back in response: they slashed the car tires and threw rocks at them [421, 434]. This protected the actors and others in the area to some extent, as it made self-driving cars less operable and sometimes forced humans inside the car to override the automated driving [434]. Additionally, this response pressured companies by making explicit the disapproval of potential customers. Similarly, when Amazon began using AI-powered cameras to surveil workers, some drivers covered the cameras with stickers so that they could no longer record, making the affected cameras inoperable [229].

In the same vein, people have put other potentially harmful algorithmic systems out of service. One way people do this involves disconnecting systems or making them otherwise unable to collect data. For example, a Carnegie Mellon University graduate student unplugged privacy-violating sensors that had been installed in his office without consent [228]. Similarly, before their exams, Chinese high-school students unplugged cameras used in facial recognition systems that monitored their behaviors during class [592]. The students had not consented and worried about accuracy, reliability, and bias within the system [592].

Depriving an algorithmic system

People sometimes refuse engagement with algorithmic systems by withholding needed inputs like their personal data or by removing those that have already been submitted. "Data strikes" in which "users withhold their data labor from a tech company, some of the company's essential services will suffer, and this would then force the company to make concessions that are desired by the public" constitute one significantly theorized about form of this response [551]. When done collectively, these responses aim to exert greater pressure but retain protective effects, whereas individually these responses protect more than they pressure.

Boycotts have been used for decades to pressure companies by refusing to engage, typically aiming to economically stunt via withholding money from companies. Boycotts operate similarly in digital contexts, often by removing user presence and thus revenue-generating ad impressions. For example, after Facebook disproportionately targeted Black users with misinformation ads, Facebook users boycotted the platform by logging out for a week [326, 561]. This put pressure on

Facebook, but it also protected users by making them unable to encounter harmful Facebook ads for a time.

People can also leave harmful systems with similar aims. In one case, African workers at Sama, which performs some of Facebook's outsourced content moderation, left the company because of poor pay, poor working conditions, and poor mental health resulting from their job functions [398]. Resigning like this mainly serves to protect people from harmful work conditions and impacts, but it also mildly pressured Sama by showing that some were unwilling to work for them under current conditions.

People can also "leave" within the context of online algorithmic systems. For instance, in her "Opt Out" project Janet Vertesi says that deleting her Google account and ceasing use of Google-related tools "aimed at keeping [her] and [her] family away from the evils of data-collecting algorithmic systems" [544]. As another example, Twitter users frustrated by algorithmic control (and new ownership) responded by leaving Twitter and, in what has been called "conscious data contribution" [550, 552], moving to Mastodon, a platform without an algorithmic feed [497]. While these examples have underlying elements of pressure, they primarily serve to create safer existences online for those responding.

People can also withhold without the high commitment of entirely quitting all use of a platform, often via some form of hiding from an algorithmic system. For example, people adopt "practices of keeping personal and private aspects of life offline, such as not sharing images of young children and babies online, using encrypted services like Signal, [...and] the adoption of anonymous social media accounts to speak to a smaller circle of confidants" [193]. People can perform more manageable data strikes with smaller actions like using private browsers or refusing to rate, review, or comment on products [551]. For instance, Bangladeshi and Indian users, bothered by targeted ads, turned off location data or changing microphone settings for some apps on their phones to allow less access to their data [471]. Less severe actions like this lead to less pressure felt by companies or other targets. Responders gain some algorithmic anonymity and freedom from data collection, but the protection is less complete than leaving a platform wholesale, with the trade-off of retaining the ability to interact with the system.

Confusing an algorithmic system

People also refuse engagement by providing false or unexpected information to harmful algorithmic systems. One version of this is "data poisoning", in which people provide "data that was created with the intention of thwarting the technology" [552]. Similarly, "obfuscation" encompasses "the deliberate addition of ambiguous, confusing, or misleading information to interfere with surveillance and data collection" [89].

This response often serves to obscure the personal data and behaviors of actors. Thus, confusing an algorithmic system can be protective. For example, Uber and Lyft drivers found their algorithmically determined ride payments inadequate, so they turned off their apps at coordinated times to feign a driver shortage and boost fares through app price surges [313, 513]. Baidu Deliveries, Ele.me, and Meituan delivery workers in China used similar methods, turning on work mode only at times or locations when the platform algorithms would not give them undesirable orders [509]. Acting individually, a user described pushing back against Facebook's advertising "data mining bot" by randomly and inaccurately liking pages [97]. Additionally, downloading and using some anti-tracking browser extensions enables this: for

example, AdNauseam clicks ads automatically in the background to confuse online trackers [264]. These actions protect people from having their data collected and used harmfully while also subtly pressuring algorithmlords by forcing them to collect inaccurate data.

Confusing algorithmic systems can protect others beyond those responding. Obfuscating data has been described as a way to reduce algorithmic discrimination [188]. And Chinese online protesters created hundreds of duplicate videos of events by mirroring, cropping, and filtering them to evade algorithmic censorship [584]. As another example, after Twitter algorithmically amplified the racist hashtag “#white-livesmatter”, Twitter-using K-Pop fans “took over the hashtag [...] drowning out white-supremacist messages with nonsensical or anti-racist posts” [283]. Besides protecting, these responses pressure algorithmlords by rendering their algorithmic systems inoperable as intended.

7.3.6 Communicating Algorithmic Harm With Others

People also respond by communicating algorithmic harm with others to *protect*, *pressure*, or *strengthen*. This communication resembles past descriptions of users raising awareness about harms [148, 474]. These responses can involve communications from disempowered people to algorithmlords, to others affected by the same algorithmic system, or broadly to the public.

“Complaining” to those with more power

When people communicate to those with more power, like algorithmlords or members of legal institutions, they typically raise issues in hopes that those with more power will take action to address the conveyed harm. In doing this, people employ their existing power to try to create safer experiences for themselves while also putting a limited amount of pressure on algorithmlords. “Complaining” can have negative, patronizing connotations; I reference Ahmed’s conception of complaints in which they can be powerful ways to change harmful structures, and I format this type of response in quotation marks to signal this [10].

Disempowered people have described submitting, flagging, or reporting issues as harmful when these options are built into platforms [148]. For example, YouTube creators harmed by demonetization errors in the advertising algorithm responded by contesting the errors using the platform’s appeals system [302]. As another example, Wikipedia users surfaced algorithmic harms to Wikipedia platform developers, collaborating so that developers could then work to address the issues [233]. In these cases, people raise issues, using means often built into an algorithmic platform or otherwise systemically supported, to employ their own powers.

Sometimes existing pathways are nonexistent or insufficient, so people communicate in other ways. For example, after Facebook unreasonably restricted some Bangladeshi users’ accounts, affected users failed to reach adequate resolution using the platform’s appeals process; in response, several users asked people they knew who worked at Facebook for help restoring their accounts [467]. And workers at Sama, mentioned previously, sent management a list of their stipulations to improve their working conditions [398]. Voicing concerns to algorithmlords exerts light pressure: collaborative convincing to make changes, or beginning what could become a more involved and hostile process of shifting power through more intense, forceful means.

To exert more pressure on algorithmlords, people voice concerns about harmful algorithmic behaviors to legislators. For example, San Diego residents, troubled by potential surveillance from a smart streetlights program, showed up at community forums and lobbied city council members [568]. And a worker at Plastic Forte in Alicante, Spain, submitted a complaint to the Spanish data protection agency (AEDP) after discovering that the company monitored employee working times using secret facial recognition [104, 120].

Publicizing algorithmic harms broadly

When communicating issues broadly, people often aim to create bad PR for algorithmlords and connect with others who might help support their cause. Thus, publicizing can be viewed as both pressuring and strengthening. These two aims feed on each other, with increased media broadcasting leading to increased public awareness and support, which in turn often leads to more media coverage.

This type of response can be as straightforward as posting about encountered algorithmic harms on social media. For example, LGBTQ+ YouTubers harmed by automated video restrictions took to Twitter to describe the issues [176, 474, 590], as did lawyer Carrie Goldberg to share widely about Amazon's harmful "suicide kits" [33]. And Weibo users posted with the hashtag "#ThankGodIGraduatedAlready" to spread awareness about surveillance cameras installed in Chinese highschools [592]. This can lead to additional support, such as through retweets on Twitter, which builds users' power. If the posts gain significant popularity, the media often picks up the issues, which pressures algorithmlords.

People also publicize algorithmic harms broadly through protests. For example, when algorithmically determined A-level exam scores affected university admissions for students, especially those from lower socio-economic backgrounds, in over 160 countries including the UK and Bangladesh, students and teachers protested and demonstrated worldwide [2, 162, 404, 472]. This garnered significant public support and many news outlets covered the protests, both strengthening those affected and pressuring algorithmlords.

People also directly contact news outlets to publicize algorithmic harms they encountered. For example, San Diego smart streetlights opposers planned a press conference to disseminate their concerns [568]. And after Johanna Burai saw only light skin in image searches for hands, she created a website and ran a media campaign to spread awareness and create more links to her site, pushing it higher in Google results [92, 542]. Directly contacting media can shift power by exerting PR pressure, and the resultant publications can help the public gain legitimacy and thus build power.

Another way people publicize algorithmic harm broadly involves writing longer, often more formal pieces or reports. This can further legitimize their own arguments to the broader public, strengthening their position, and provide resources external to the system for others. Additionally, these writings exert pressure when they are shared and spread widely. For example, the smart streetlights opposers created a policy report, which they also publicized by sharing its findings with journalists [568]. And after researcher Joy Buolamwini's noticed facial recognition could only detect her face with a white mask, she wrote and published a research paper about the issues [91, 321].

Sharing, discussing, & organizing with others affected by the algorithmic system

People also respond by sharing information with others affected by the same algorithmic system. This communication can aid in the formation of counterpublic spaces, or “parallel discursive arenas” [186], like those in which users collaborate and discuss potentially harmful algorithmic behaviors [474]. These spaces help similarly affected people strengthen their positions, exchanging useful knowledge and building power in numbers, and protect themselves via the development of solidarity, which can actively make algorithmic spaces safer for them.

Sometimes people simply convey information to others. For example, when Booking.com users noticed the rating algorithm calculated higher overall ratings for hotels than they intended, they warned others about this in the text portion of their reviews [168, 474]. As another example, TikTok users saw harmful content on their For You Pages and created videos of their own describing the problematic elements of that content to others [485]. In these ways, users alerted other users to harms present and tried to educate those who might not recognize the same harms.

Beyond simply sharing information, people also discuss issues with others affected. For example, Chinese riders for food delivery platforms created and joined small WeChat groups to develop networks for solidarity, sharing knowledge like difficult delivery locations [591]. And African workers at “ethical AI” company Sama created and used a WhatsApp group to discuss traumas of moderation work, low pay, lack of benefits, and unreasonable work hours [77, 398]. In these cases, people developed mutual support to protect themselves while discussing the nuances of an issue and potential directions for remediation, which helps them build power.

Finally, people also organize into collectives to take action with others, which can be used to protect themselves and pressure algorithmic lords. For instance, smart streetlight adversaries in San Diego formed a coalition called TRUST San Diego to work together to counter the smart streetlights program [568]. The Chinese food delivery workers mentioned above used their chat groups for mutual aid amongst themselves, supporting each other in ways like informal transfers of orders to get deliveries completed on time and the formation of unofficial unions that provide increased bargaining power [77, 509, 591]. And African content moderation workers for ChatGPT, TikTok, and Facebook formed the African Content Moderators Union, strengthening themselves to push for better pay and work conditions [397].

7.4 Limitations

Though I attempted to ensure that the responses in our taxonomy are drawn from cases around the world, an early overrepresentation of Global North cases suggests a bias in the field that impacts our work despite our efforts. Regardless, differences between observed cultures and their norms of behavior, between designers’ values and the relevant actors’ values, and between pertinent legal systems can drastically affect how extensible a response is from one context to another. For example, a response of public protest that works well in one place might be much more dangerous or less efficacious in another. Additionally, I did not try to collect undocumented responses, so our analysis and taxonomy tends toward types of responses that receive some amount of visibility or leave some trace. Future work should explore responses from below in additional contexts and cultures, as well as in other languages and formats. Further, while our broad scope allowed development of an overarching understanding of responses from below, it necessarily abstracted away many specifics of individual cases, such as different actors’ affiliations, incentives,

possible actions, and complex and diverse relationships with power like their positions in the matrix of domination [253].

7.5 Discussion

I have presented ways that people currently respond from below to algorithmic harm, through a taxonomy that connects different types of responses to existing theorizations of power. In this section, I discuss significant challenges related to supporting responses from below and highlight open opportunity spaces for future research to better support disempowered actors and their responses to algorithmic harm.

7.5.1 Burdens of Responding to Algorithmic Harms From Below

While our work has shown the potential and power in driving toward positive change from below, it is critical to consider the burdens this can have on the actors, particularly members of marginalized communities. In this work, I observed that documented responses to algorithmic harms largely come from communities affected by algorithmic harm (e.g., people of color, sexual and gender minorities). Thus more burdens are thrust onto members of marginalized and underserved groups who are already more prone to be exposed to and negatively affected by societal and algorithmic harms [141].

These burdens can further disempower people who respond from below. When helping other actors like industry practitioners to address algorithmic harm, their efforts tend to recede from view. In addition, many types of responses burden responders with the work of transforming and transporting knowledge between cultures, known as translation work, of some kind [374]. For instance, translation work can aim to convey the legitimacy of information about algorithmic harms to those with greater power [508]. And efforts to support the disempowered in responding to algorithmic harm can simultaneously serve to prop up existing systems of power. For example, Turkopticon was intended to support Amazon Mechanical Turk workers, but its public legacy emphasized the value of design rather than the value of labor or laborers [270].

Ideally everyday people's concerns would be valued by default, but given the reality I see translation work as invaluable, especially when governmental institutions and actors from below are not in sync and communication in comprehensible formats is crucial for alignment. The FAccT community can collaborate with everyday people affected by algorithmic systems to build tools designed to push algorithmic complaints into view in ways that cannot be ignored. Building on existing repositories that collect and document real-world algorithmic harms (e.g., [4, 5]), platforms like this could be developed with and for everyday people, focusing on them as the primary users and contributors, and pressure algorithmlords by bringing issues to the attention of policymakers and the broader public.

To lessen the burdens of translation work on everyday people, I see opportunities for the FAccT community both to help actors from below most effectively frame findings from specific investigations and to proactively shape what forms of knowledge and evidence are understood as legitimate by powerful actors, such as algorithmlords and policymakers. The FAccT community has opportunities to elevate people's complaints to ensure that they are heard and acted on, recasting what affected people have already been saying into forms that will be appropriately valued

by those with more power [374, 508]. In this vein, Ahmed breaks down the labor behind making complaints legible to an institution [10].

That being said, affected people should not have the burden of duty to surface and address issues as they arise; ideally, algorithmlords should preemptively address algorithmic harms so that affected people are not forced to take action. Changes after harm has occurred can be insufficient: replacing algorithmically generated A-level scores after protests, as described in Section 7.3.6, still left an “algorithmic imprint” of different scores than had the algorithm never existed [162]. With policymakers, the FAccT community could support actors from below by developing new systems and structures that incentivize algorithmlords to proactively identify, recognize, and address issues, and especially holding them accountable to harms they put into production [552]. For example, the US Federal Trade Commission recently penalized a company with a harmful algorithmic system, forcing them to pay a fine and delete data and related AI models [82, 295].

7.5.2 (Adverse) Impacts of Responses From Below to Algorithmic Harm

Responses from below can significantly help remediate algorithmic harms, but they can also have less desirable effects. For instance, broadly publicizing Amazon’s harmful “suicide kit” recommendations [33] also makes more people aware of this suicide method. Such consequences become even more challenging with the presence of bad actors who misuse responses, intentionally or not, to algorithmic harms.

It is difficult to fully anticipate what sorts of impacts a response from below might have down the line, especially given how expansive they can be. In this research I saw these responses have both individual and more widespread impacts. Additionally, although I focused on more immediate effects of responses in this chapter, responses can also have significant indirect effects.

While there might not be well-defined ways to identify and eliminate the adverse impacts, designers and developers who aim to empower responses from below to algorithmic harms can make harmful misuse more difficult. For example, Fawkes is a researcher-developed system that allows users to protect their digital images from unwanted facial recognition by making visually imperceptible pixel-level changes [470]. In theory, someone could use Fawkes to impersonate someone else. But because Fawkes was designed for avoiding identification in *unauthorized* models, for which most people want to avoid being recognized, being mistakenly identified contributes to the collective goal of obfuscation without causing harm to individuals.. Additionally, to further retain focus on mitigation of algorithmic harms, designers could create tools aligned with principles of restorative and transformative justice, making them more difficult to redirect toward harm [444, 582].

The FAccT community might also further consider concepts of consent and agency. I see opportunities for researchers to investigate, understand, and work toward a world in which consent and refusal are equally viable options [58, 484]. Zong and Matias highlight the ways in which refusal can preemptively avert harm [596]; to that end, the FAccT community should also consider their own agency and when it might be most beneficial to not design, to not deploy, or to declare insufficient and retract [53, 143, 258].

7.5.3 Obstacles Faced When Responding From Below to Algorithmic Harms

The specific contexts of algorithmic systems present unique challenges for responses from below to succeed. It can be challenging to pin down issues and productive directions forward due to the opaque nature of algorithmic systems and the emergent nature of algorithmic behavior in real-world use contexts [96, 190]. Additionally, it can be unclear whom to hold accountable and whom to contact for recourse, due to the diffuse and debated nature of responsibility for algorithmic behaviors [322]. Furthermore, different groups of people have different values around algorithmic systems [274], which can impact whether and how they respond. For some, the dynamics of power around these systems create a view of AI as necessarily authoritative and correct, which can lead them to take on responsibility for system harms themselves and to abstain from responding to address these harms [287, 420]. And finally, for some, entire categories of responses may be infeasible. For example, one response I observed involves leaving an algorithmic system—which is infeasible for some, such as rideshare drivers whose livelihoods are attached to the system. As highlighted by such cases, people often lack the power to fully disengage from algorithmic systems [252].

Algorithmic contract. Similar to theories of a social contract [333, 436], the public has found themselves within what I call an “*algorithmic contract*” in which they have theoretically consented and chosen to trade information about themselves and their behaviors in exchange for the computational promise and benefits of algorithmic systems. Although this algorithmic contract purports to be an agreement between everyone, in reality it is difficult for the public to disengage from algorithmic systems; more realistically, the algorithmic contract is an agreement between technological giants, in which algorithmlords with power to determine which algorithmic systems to use, how to use them, and where to deploy them have consented to these algorithmic systems for the rest of society. This parallels ideas of a technocracy, in which experts are the ruling power [173, 207], and leads to the current situation, one rife with algorithmic harms [442]. When people attempt to take back some control over the algorithmic systems that affect their lives via responding from below to algorithmic harms, they do so with few legal rights.

The FAccT community can work to improve legal protections and rights, and thus power, of those responding from below to algorithmic harm. I see researchers as having a responsibility to intervene and participate in the development of policy, helping create needed regulation that can have concrete positive impact in specific, real-world situations [493]. For instance, researchers could push to extend rulings like those recently won by Sandvig and the ACLU, which increased protections for researchers and journalists who break website terms of service to audit potentially harmful systems [496], to cover the public and their investigations. Researchers can also work with policymakers to create regulations that help people leave harmful systems. For instance, the EU’s General Data Protection Regulation (GDPR) requires that users consent to data collection and stipulates users’ rights to have their personal information erased [423]. Laws like this start to make it possible for people to fully separate from algorithmic systems and are especially important to actors for whom leaving an algorithmic system might be riskier. Additionally, researchers could develop and lead more workshops, special interest groups, or other discussions (e.g., [318, 479, 585]) that focus on ways that policy can support people’s responses to algorithmic harm or render the responses unnecessary in the first place. FAccT community members can also serve as expert advisors providing oversight and nuance for legal pathways that shift power to help ensure that they sufficiently

remediate algorithmic harm.

I see opportunities for the FAccT community to reframe the idea that algorithm-lords or others with systemic power and expertise retain control throughout the deployment of an algorithmic system. The development of new forms of governance of algorithmic systems could allow for everyday people to be involved and empowered from the start, *by design*.

7.6 Conclusion

In this chapter, I characterize real-world cases of responses from below to algorithmic harm and synthesize these cases into a taxonomy that categorizes responses based on their behaviors and their power-related aims. In doing so, our taxonomy offers the FAccT community (1) a set of major types of responses from below that future work can aim to support; and (2) a synthesis of existing theories on power and actions from below in and around algorithmic systems from fields such as anthropology, HCI, and communication. I call on FAccT community members to actively work with affected people to empower their efforts from below to safeguard themselves from algorithmic harm, and I present several existing opportunities for this.

Chapter 8

Algorithmic Users' Conceptions of Power

This chapter is adapted from my paper:

Alicia DeVrio, Jessica Hammer, Kenneth Holstein. “‘Wow, we can’t do shit’: Opportunities to Support Social Media Users’ Conceptions of Power With Games”. In Prep.

Recent HCI research highlights how, in response to harmful and otherwise undesirable algorithmic systems and insufficient changes from those in control of algorithmic systems, users of algorithmic systems like social media platforms can be powerful actors with significant abilities to influence and shape the algorithmic systems they operate in. But how do users understand their own power to influence algorithmic systems? I investigate this question in the context of algorithmic social media, drawing upon research that foregrounds how games can support understandings of power dynamics in complex systems. Through 16 interviews, I find that users perceive their power as limited and individualistic, that sharing support facilitates user connection and control, and that feelings of safety can enhance user agency. From these findings, I identify opportunities and discuss directions for future game design to help algorithmic social media users with systems thinking, collective action, and exploration and play such that they might more fully realize their power and take action using it.

8.1 Introduction

Recent HCI work has shown that in response to harmful and otherwise undesirable algorithmic behaviors that those in charge of algorithmic systems fail to appropriately remediate, users of algorithmic systems like social media platforms can be powerful actors with significant abilities to influence and shape these systems (e.g., [148, 149, 313, 474, 552, 568]). For example, in data strikes, users collectively remove their data from a platform to starve its systems and gain leverage themselves [551, 552], and in “everyday algorithm audits”, users work together to test potentially harmful algorithmic systems and to raise awareness about issues in order to push for change [474]. In these examples, users’ collective power can successfully effect change by pushing those in charge of algorithmic systems to more sufficiently address issues than they likely would have otherwise [149]. To that end, existing research has emphasized how, similar to workers organizing into unions to agitate for

better conditions, users can be understood as laborers with abilities to exert worker power to improve their algorithmic experiences via collective action [37, 325].

But to what extent do social media users view themselves as powerful actors? In this work, we asked **RQ1: How do social media users conceptualize their power to influence social media platforms?**

It can be challenging for people to fully grasp the complex dynamics of power around systems they operate in. Significant research highlights the advantages of games for helping people understand complex systems and causal relationships. For instance, one strength of games is their ability to model systems and their dynamics [79]. Additionally, game players can interact with and shape these systems (e.g., [238]). Thus, we asked **RQ2: What opportunities exist to support social media users' understandings of their own power on social media platforms through the use of games?** By understanding users' current conceptions, we can better understand how games might effectively intervene to support their understandings of social algorithmic systems and their own powers within these systems, empowering them to take action to shape these systems in ways they want and need.

To address these questions, we conducted 16 semi-structured interviews with users of algorithmic social media platforms who also play games. We investigated how social media users understand their own power and capabilities on social media platforms, finding that users perceive their power as limited and individualistic, that sharing support facilitates user connection and control, and that feelings of safety can enhance user agency. Based on these findings, we identified opportunities for games to help algorithmic social media users with systems thinking, collective action, and exploration and play; as such, we discuss directions for the design of future games that can support users in more fully realizing and taking action using their power.

8.2 Related Work

Here we provide a brief overview and discussion of existing work related to algorithmic systems and games. Specifically, we detail research on power dynamics in algorithmic systems as well as users' power in algorithmic systems, and highlight some related games and games literature for each of these.

8.2.1 Power dynamics & games

Algorithmic systems often produce less-than-ideal experiences for users of those systems. For example, there is significant HCI research on the harmful behaviors produced and perpetuated by algorithmic systems (e.g., [152, 379, 459, 473]). A growing body of HCI work has highlighted how algorithmic harm and other issues stem from structural power differences (e.g., [22, 69, 74, 149, 284, 299, 364]). For example, Alkhatib describes how algorithmic systems "specifically misread marginalized groups disempowered by the structures of power surrounding the systems that cause harm" [22].

Everyday users are often unable to rely on those in control of algorithmic systems to shape these systems in ways that are suitable and beneficial to users [149]. For example, Facebook not only unfairly restricted the accounts of multiple Bangladeshi users but also failed to provide satisfactory contestation mechanisms for these users, leaving them to handle their restrictions on their own [467]. Everyday users are

rarely brought into the design and development process of algorithmic systems—and even when they can participate, they rarely have significant power over decisions [123, 487].

Games research showcases the significant potential for games and play to help people learn about systems. For one, games and play have long been studied as sites to help people learn (e.g., [187, 197, 304, 478]). Additionally, games' abilities to model systems and their dynamics have been described as "the greatest strength of the medium" [79]. And players have abilities to interact with and shape these systems: for instance, Hammer et al. describe how in role-playing games, players use their power relations to change the game world, get their way when others disagree, and define what is valued [238].

There are many games and playful educational experiences focused on algorithmic systems. For example, *most likely machine* has players build their own algorithm and watch its consequences [38]. *Interland* teaches players how to be kind and protect themselves from other users online [215]. *Bad News* and *Choose Your Own Fake News* teach players about mis- and dis-information [272, 435]. And in *Can you break the algorithm?* players are journalists investigating an algorithm with harmful impacts [19]. Games like these broadly aim to support players in learning about algorithms and how they work.

However, few games to date focus on the power structures of algorithmic systems. Some focus on the role of tech companies and their choices: *The Future of Memory* is a text-based game in which players are workers at a social media company who ultimately participate in censoring online speech [558], and *Trust and Safety Tycoon* has players manage a trust and safety team and navigate the choices that come up [353]. *#ForYou: A Game About Algorithms* has players act as social media content creators who want their content seen more broadly or as a social media company that wants to make more money [488]. Similarly, in *Unveiling Interfaces*, players publish apps in an algorithmically-driven marketplace [384]. Both these games focus on helping players understand how algorithms work, though they do also allow for some critical engagement of users' abilities within systems. However, they conceive of users' roles as primarily individuals who must take whatever the algorithm doles out.

8.2.2 Users' power & games

Disempowered users of algorithmic systems have taken action themselves to shape algorithmic systems in the ways that they want and need, through forms of their own "everyday resistance" [460] and "tactics from below" [568]. DeVrio et al. collected and taxonomized many ways that everyday people respond to algorithmic harms [149]. For example, cases of everyday algorithm auditing illustrate how everyday users interrogate, collect evidence about, and push for remediation of algorithmic harms they encounter in everyday contexts [474]. And cases of data leverage like data poisoning show how everyday people can provide false information in attempts to reduce a harmful system's ability to function [552]. In a more labor-focused context, Kellogg et al. describe strategies workers use to resist algorithmic control, which they term "algoactivism" [296]. Other work has proposed that users' data (or simply engagement [276, 518]) can inform an understanding of users as laborers, which in turn gives rise to ideas of how users could leverage power as workers that largely comes via forms of collective action [37, 325, 405].

Actions like those described above frequently involve groups of users uniting and using collective power to push for change. For instance, data leverage and everyday algorithm auditing are conceived as being powerful enough to exert change particularly when many are involved. Similarly, the book *Algorithms of Resistance: The Everyday Fight against Platform Power* puts forth that “computational power can be confronted only through cooperation, mutual aid, and collective power” [76]. Carmi and Nakou also conclude that “collective mobilisation [is] necessary to intervene in existing institutional hierarchies and achieve a significant social and political change” [103].

That being said, collective action can be challenging to achieve, especially in algorithmic contexts. For instance, Viljoen discusses how “datafication does not result in the kinds of visceral oppression that spurs organizing and builds counter-power” and goes on to describe challenges around forming collectives and working together to push for change [547]. Similarly, Li et al. caution that data labor that does not require connections with others “makes it difficult to organize collectively and gain power to influence technology operators” [325].

At the same time, there are many games focused on collective action in labor contexts. These tend to be more experiential, directly relating play to power structures and labor relations. For example, *Trapped*, *Against the Grain*, and *The Price of Coal* are all set during specific historical events—the 1911 Triangle Shirtwaist Factory fire, a 1944 World War II production company, and the 1921 West Virginia coal miners’ Battle of Blair Mountain [45, 526, 567]—that allow players to experience real-world contexts. Other games focus on labor relations more detached from specific labor contexts, such as the cooperative game *STRIKE! The Game of Worker Rebellion* and role-playing game *Comrades* [17, 282]. In *Bitches in the Club*, players role-play as exploited service workers who want to regain control of their club [533]. And *’Twas the Labour Dispute Before Christmas* and *Beat the Boss* aim for people to learn specific union organizing skills while playing [202, 495].

At the same time, we identified just one game that delved into collective power on social media: *Kristine Is Not Well* is a VR simulation of resistance to algorithmic surveillance and censorship [409]. This dearth suggests opportunities for games and playful experiences that allow players to interact with, examine, and better understand their powers within algorithmic systems, helping them learn about how they might effect change within the complex dimensions of algorithmic power structures.

As such, we focus in this work on understanding the ways that users of algorithmic systems see their own powers and how games might support these understandings. In line with recent HCI work on empowerment, we especially aim to better understand how to help the disempowered to confront those in power [456]. We orient our work toward understanding opportunities for everyday people to fully understand and take action with their power.

8.3 Methods

In this study we set out to investigate how algorithmic social media users understand their own power on social media platforms, as well as what opportunities might exist for further support of these conceptions by using games. To explore this, we conducted 16 semi-structured interviews with adult social media users. We use P1–P16 to refer to the participants, and all participant quotes in this chapter are from these interviews. Our institutional IRB approved this study.

We recruited participants via social media (i.e., Twitter, Facebook, Instagram, NextDoor) and Craigslist Jobs in August 2023. As we wanted our interviews to cover a variety of algorithmic social media platforms and types of people, we directed interested people to fill out a brief screening survey to provide information about what social media platforms they had an account on, which social media platform or platforms they used most often, and basic demographic information (age, gender, race/ethnicity, education). Following [148] we framed our demographic questions based on the US Census survey [93] and other HCI research [455]. We received 143 responses to this form, from which we selected participants who spanned a diverse range of algorithmic social media platforms and demographic backgrounds.

Our recruitment process provided a set of 16 interviewees (69% women, 12% men, 25% nonbinary—participants could select multiple genders) ranging from 19 to 71 years old with an average age of 39. Two of our interviewees were transgender, 12 were not, and 2 did not wish to say. Our participants were 62% white, 6% Hispanic or Latino, 19% Asian, 6% Middle Eastern, and 25% Black or African American. (Participants could select multiple races/ethnicities.) For 19% the highest level of school they completed was high school, for 25% some college, for 12% a bachelor's degree, for 19% some graduate school, for 19% a master's or professional degree, and for 6% a doctoral degree.

The semi-structured interviews took place in August and September 2023 remotely over Zoom. Each interview lasted 30–60 minutes, and participants were compensated with \$20 Visa giftcards. Social media platforms' algorithmic systems can differ, so we centered the questions in each interview around experiences of a specific popular algorithmically driven social media platform, based on the platform the participant reported that they used the most in the screening survey: 6 interviews focused on Facebook, 3 on Twitter, 3 on Instagram, 2 on YouTube, 1 on LinkedIn, and 1 on Reddit. As we are interested in broader, non-platform-specific implications regarding user understandings and actions related to their own power, we focused our questions on this: that is, to structure our interviews, we iteratively wrote, piloted, and revised a set of questions that cover 1) thoughts about and behaviors on algorithmic social media, with a focus on conceptualizations of user power and related actions and 2) experiences of and feelings about playing games, with a focus on power. We recognize that significant work exists on experiences of and feelings about power dynamics within games; in our interviews, we ask questions about these areas not to generate novel findings about games but to collect responses about both algorithmic social media platforms and games from the same participants, allowing for fuller interpretation and more direct comparison in our analysis.

Using transcripts produced from recordings of the interviews, we conducted a bottom-up thematic analysis of the data from the interviews [119]. As described above, we aimed to understand broader implications about how social media users conceive of and act with their own power, so we conducted our analysis by especially looking for patterns across our participants and across the different algorithmic social media platforms in our interpretation of results. The first author manually thematically coded all interview transcripts using ATLAS.ti, aiming to capture a wide and varied range of qualitative observations about how participants conceptualize and understand their power on social media platforms, the actions they take with their power on social media platforms, what factors might affect their conceptualizations and actions, and how their conceptualizations and actions on social media platforms compare to their conceptualizations and actions in games contexts. The full research team then engaged in a series of interpretation sessions to discuss and

identify important observations about codes as well as to cluster codes into higher-level themes. We present these themes below.

8.4 Results

In this section, we explore how 1) participants understand their power on algorithmic social media platforms and 2) how these understandings compare with their understandings of power within games. Based on these explorations, we identify three opportunities for games that can support social media users in more fully understanding their power and taking action in algorithmic social media contexts. In the following discussion section (§8.6), we elaborate on these opportunities, outlining specific directions for the future game design.

8.4.1 Users perceive their power as limited, individualistic: The need for systems thinking

Participants mostly described their power on algorithmic social media platforms in individualistic terms, suggesting opportunities exist for social media users to develop greater understanding of the system-level impacts of their actions. In response to questions about how much control they had to change the way social media algorithmic systems work, aside from P4 who described observing but not participating in collective action, all participants either said they did not have much power or described actions they took on their own to try to change the way the algorithmic system behaves. For example, participants described trying to control algorithmic behavior by *“changing the type of photos that I post [...] if I’m trying to play with the algorithm”* (P12) and *“reporting spam or harassment”* (P1). In addition to describing individual actions, participants tended to describe the impacts of these actions at an individual level: they described their own powers in terms of how they could influence the behaviors of social algorithmic systems for themselves—that is, how they could affect their own, personal experiences. For instance, participants described how their engagement actions could be powerful inputs that contributed to shaping their own algorithmic experiences on social media platforms. P10 said, *“Let’s say I join a cat subreddit. And I think Reddit sees this, the algorithms are like, ‘Oh, she likes cats, so let’s suggest to her other similar cat subreddits.’”* In this example, P10 identifies the content shown to them as a direct result from their earlier actions. Participants also saw their own actions as affecting the ordering of content they are shown: *“Things that I click on more frequently would come up first”* (P2). Participants extended this understanding to purposefully leverage their interactions to shape their experiences. For instance, P11 said, *“I wanted to eat more plant based, so I went through and clicked on a couple of content creators that were showing me recipes like that, and then it was like, that’s all I saw for a while.”*

Games offer ways for users to recognize the potential system-level impacts of their actions on algorithmic social media platforms. Games are known to be effective in helping people understand system dynamics and how actions have impacts within those systems beyond the individual (e.g., [79]). Participants frequently described their impact in games as broad, talking about their abilities to control major elements of game worlds. In role-playing games, participants expressed feeling *“control over my environment and the choices that I’m making”* (P6). For instance, P12 felt powerful from *“just being able to move through the space and be in charge of like, what*

tool are we going to use. [...] I wasn't told that there was only one choice to me and I had to follow it."

Additionally, participants expressed feeling little agency on social media platforms, describing feeling their actions did not matter or they could not take much action on social media. This signals opportunities to scaffold where users could act as well as to bring more visibility for users to see the impacts of their actions. For example, P4 recounted a time when user actions seemed not to matter: *"a lot of the subreddits kind of closed down in protest"* to a change in Reddit's API but that *"most of those [subreddits] are back up right now [...] and it definitely is a release valve for how people are feeling, but it often doesn't do anything."* P4 concluded, *"I can't really think of many moments where I'm like, 'We do have power.' It's mostly moments in reverse where I'm like, 'Wow, we can't do shit.'" Similarly, P15 described feeling stuck, out of options for action: "I've tried really every which way [...] there is really no other way to get [Instagram's algorithms] to change much honestly."* In contrast, games are often specifically designed to enhance feelings of agency, such as with flexible, dynamic stories that can allow a freedom of choice that enables players to feel their influence. For example, P4 described how in computer role-playing games like *Skyrim*, *"You can run around and be whatever you like. You could be an assassin. You could be a maid. You can be the head of the dark brotherhood. [...] You could be a werewolf."*

Participants identified corporate individuals as in charge of social media, suggesting opportunities to support participants' understandings of broader structural power dynamics, both within and beyond powerful institutions like social media companies. For example, participants described *"Mark Zuckerberg, CEO"* (P3), *"Elon Musk [...] and the rest of the exec board"* (P7), *"a Facebook headquarters with a bunch of scientists"* (P11), and *"the 10 dudes in power in general in the world"* (P12) as in charge. This suggests that some participants tried to make sense of power and systems using the kinds of reasoning and attributions that make more sense for individuals (e.g., see [237, 349, 527]). Games offer a good potential way to help social media users broaden their understandings of power, as games are well equipped for showing systems of power. Participants occasionally described structural power dynamics in which institutions like social media govern them. For instance, P12 continued to describe how *"power begets power, money begets money [...] most social media is just the latest institution that is kind of governing us"*. This showcases participant understandings of the disperse nature of power in social algorithmic systems (e.g., [572]) and highlights an opportunity for participants to understand their own roles within broader power structures, to recognize that they have power within these systems. Games are well suited for helping players see their roles within systems of power and their potential to impact these systems of power.

Participants did see themselves as providing value to corporate entities, but they saw this value as minimal, especially small when compared to the total production value of a social media platform. This suggests participants needed better understandings of the collective value of users on social media platforms, as well as how this value is central to social media—that is, social media could not exist without users. For instance, expressing how users are valuable to companies, but mainly for the purpose of data gathering, P4 said that users provide companies with *"information that you're kind of freely giving them in return, so like your preferences, what people you are interested in, what you buy, where you live, [your] social life, your literal social networks, like the people you know."* A couple participants recounted how users can make money from the content they contribute to a platform, but only if the platform is also making money from those users. For example, P12 said that Instagram *"let you monetize reels if you have a certain number of followers, so every reel that you posted for a month, you*

could earn up to \$1,200. \$1,200 was for 11 million views [...] you have to wonder how much [the platform is] pulling already from those reels." Similarly, P9 said, "There's ads, and so people who make videos make money, but also YouTube itself makes money from all the ads that it hosts.". So while participants saw social media users as having some value, they understood this value as dwarfed by the value generated by social media platforms. Games have potential to support these conceptions by helping players reflect on their individual and collective value within game contexts.

8.4.2 Sharing support facilitates connection and control: The need for collective action

In line with previous literature (e.g., [28, 277, 464]), participants saw connecting with others and forming community as a strong motivation for their use of social media platforms, and they valued their connections for how they facilitated the sharing of support. This suggests that helping users continue and enhance their sharing of support on social media can help them form connections, which are central to collective user power on social media. For example, P15 expressed how they feel supported simply from observing their connections on social media: *"if you're feeling a little lonely or down [...] it's just kind of a nice way to see what everybody else is doing."* Participants also described ways they intentionally drew on their connections for acts of support. P2 described using a Facebook group to *"literally chat every day now so that it's like we're a support group."* And while stuck in an airport with back pain, P13 said, *"I was posting [on Instagram...] and then when I came back [home] almost everybody checked in with me to see if I'm safe or not."* In these examples, social media connections directly supported participants through messages showing care.

Similarly, participants valued connections in games for how they facilitated the sharing of support. Participants described how playing games with others could be supportive. P13 explained how support can occur in games from mutual recognition: *"I see you, you see me [...] just like liv- being there with each other."* And P3 described how player connections can support learning real-world skills: *"My son, he needs help with his social skills. I play games with him, but he also plays with his friends, and then he feels more confident, overcoming his shyness."* Player connections can extend beyond the realm of gameplay itself as well. For example, P6 discussed getting to know players outside of gameplay and leaning on each other in a variety of ways, from *"asking somebody for advice about [...] looking to buy a new laptop"* to *"somebody is in high school calculus and needs help on their math homework."* This example highlights how support from games can extend beyond the interactions explicitly supported or anticipated in gameplay.

Within the realm of support, participants described taking actions to impact others' experiences on social media, in contrast to how they conceived of their influence on social media as more limited in scope, as described earlier. This suggests opportunities to help users understand that the kinds of actions they see as supportive are the same actions they can take to have wider influence on social media—that is, there are opportunities for users to learn to take support actions not only in the realm of support, or to expand their notions of the realm of support. For example, participants described supporting other users via intentional actions to manipulate how a platform's algorithm behaved. P6 said, *"One of my friends who is a tattoo artist will post that they're opening commissions or post art or recent work that they've done, and I always will like and maybe even try to comment on those [...] 'cause I'm aware that that engagement also helps boost their content to other people."* In this example, P6 chose behaviors with the explicit aim of impacting other users' experiences. Participants also

described seeking and providing support in ways that typically involve acting more collaboratively. For instance, P3 said that people in need *"might go to a mutual aid group"* in which users could work together to share resources for support. And P6 said that social media is *"useful to amplify"* things like *"activist groups and news about like mutual aid projects."* This supportive amplification requires many users to boost information in tandem.

Games can help users see shared incentives, helping foreground ways that support and change can be accomplished by acting together. Impetus for collective action is especially needed because, though participants described acting together for support, when talking about acting together to change how social media works, they described themselves passively, more as observers than organizers. For example, P10 described a group of Reddit users who came together to create and amplify information raising awareness about a problematic company running an ad campaign on the platform. However, about their own involvement, P10 said, *"I'm not very—I would like to be more active and be a part of these campaigns, but I don't think I've ever really consumed social media quite like that."* In the same vein, P4 described how sometimes posts that involve *"bullying"* people in charge of social media can encourage those in charge to *"reverse a horrible decision"* about the platform, but that they do not post on social media themselves. These descriptions signal that participants are aware of how they might engage in collective actions in attempts to change how an algorithmic system works—in addition to their already existing participation in collective acts framed as supporting others—but that they need an additional push to take action themselves to push for desired changes.

In contrast, participants frequently described ways they collaborated with others while playing games, working collectively toward shared goals. For instance, P2 and P9 described working together with others to brainstorm and help with gameplay. About this, P6 said, *"I think that [it] is like, morally or even psychologically or spiritually important in some sense to experience being on a team, communicating with each other."* And P10 emphasized how collective experiences were linked to the controls afforded by games, explaining ways a game highlighted *"how our decision-making, how we interact with other people, how our choices affect others, both now in the immediate and in the long term."* As such, games offer ways to foster players' participation in and even initiation of acting collectively via incentivizing and providing avenues for collaboration during gameplay.

8.4.3 Feelings of safety enhance agency: The need for play & exploration

Participants described how knowledge of rules and system logics impacted their feelings of control, highlighting opportunities to enhance social media users' feelings of and abilities to act to control by helping them have more knowledge of system rules and logics. In some examples, participants recounted times when they were able to take actions that resulted in desired, planned outcomes in games and on social media, showcasing how knowledge of rules and system logics can allow people to anticipate the influence they might have and feel some level of control. For instance, participants recounted understanding and working with the logics of algorithmic social media systems: P15 described *"utilizing more of the features that I wasn't before"* to add *"pizzazz"* such as tagging the location and other people in posts to get more views. Similarly, in a games context, P6 said, *"I tweak one more thing where I find the right setting, and then I'm able to beat a difficulty level that I wasn't able to beat before."* In other examples, participants recalled times they were unable to change or directly inspect the system rules under which they operated, highlighting how

unclear or missing rules and system logics, whether in games or on social media, can lead people to feel they have little control. For instance, P7 described cryptic requirements for reporting harmful content on social media: *"Twitter either [...] says this doesn't break any of the guidelines or like this is a fine, correct Tweet, which I would disagree with."* And P12 described these confusing requirements as a poster whose content was repeatedly removed from Instagram: *"Every time it happens it truly surprises me [...] Instagram tells you a set of rules and then applies them in a way that is like totally all over the place."* In games contexts, participants also described being unable to change or directly inspect system logics: P6 said they *"really wish there would be some ability to permanently just put somebody on a block list,"* and P1 talked about a card game with unknown rules that required *"a lot of trial and error. Like you'll play a card or a sequence of cards, and they'll be like, 'no this is wrong.'"*

To that end, some participants expressed interest in having greater control over granular adjustments of social media algorithmic systems, highlighting opportunities for games to help people accomplish this. Greater control on social media could involve, for instance, that *"people come up with their own bespoke personal algorithms"* (P6). Participants described similar desires fulfilled within certain game spaces that allow people to experience and play with granular control. For example, P15 said about a game, *"it's just cool to be able to customize or do whatever you want. [...] I love to be able to pick my character and pick everything on the car down to like mirrors and exhaust or whatever, or being able to customize tracks or customize the music, or like being able to curate your own experience and make it yours."* This suggests games could help social media users experience and accomplish control.

That said, many participants did not want to burden themselves with having more algorithmic control on social media, pointing to a need to shift perceptions of control from being seen as work to being seen as opportunities, even fun ones. For example, P14 said, *"I don't necessarily want to shape the content that I see. [...] I don't think I'd want to be bothered with making lots of decisions all the time."* Additionally, P5 worried that with more controls, *"I would get carried away with nerdiness, and then I would lose the whole point of being on Facebook, which is the fun of it."* Essentially, participants worried that more controls on social media might involve work, and that this work of control might diminish their enjoyment of social media.

Games, however, offer potential ways to help people feel more excited about possibilities for enthusiasm and joy with control rather than just staving off the negative in a deficit-focused way. In contrast to their feelings about control on social media, in games participants felt curious about involving themselves in complex decisions. P10 was intrigued by the breadth and intricacies of role-playing games: *"So you have that curiosity about, what's this world like? What are the neat things about this world that make it different from my own world? What are these characters like?"* And participants described being interested in and drawn to games involving puzzles and complex thinking. For example, P11 said, *"I used to love that [game] because it was hard [...] and it was challenging."* Similarly, P15 said, *"Ooh I get to figure out this puzzle [...] and I kinda use that time to have fun with the journey rather than trying to move through as fast as I can."* Perhaps it is unsurprising that people might be more eager for control when it feels closely connected to enjoyable pursuits rather than staving off negative experiences.

Speaking of negative experiences, participants described vulnerability, little control, and reluctance in response to bad behavior they encountered on social media, suggesting a need to change these feelings so that they can feel and exert their power against this bad behavior. Participants talked about how they do not know how to control situations on social media where others *"can put information out there about*

people that they don't want out there, like not getting consent" (P2) and where "somebody could post something where they're talking about their experiences, and seeing how quickly people would attack them for those experiences. [...] That's part of the reason why I'm reluctant to actually create my own content" (P10). Situations like these can translate to feeling threatened.

On the other hand, in games participants often saw bad behavior as a challenge rather than a threat. For instance, P2 described having influence over game characters: "You have to protect different characters from bullies. [...] You're helping these other people, and you defeated a bully. So that's powerful 'cause you defeated a bully, powerful because you made someone who was dejected feel happy." And P10 described fighting wicked characters in games as "A dopamine hit! [...] When you are just running around killing all the monsters, it's like you get a nice, huge dopamine fix because it's fun [...] and you're getting stronger." Feeling challenged in these situations might arise in part from knowing that success is possible, that there exists some way to overcome difficult encounters by design of the game.

These distinct feelings of control also had bearing on participants' feelings of safety and vulnerability in games. For example, P6 said about games, "There's a level of safety. I feel like I'm in an environment that I understand, that I sort of have control over, which is nice when you don't necessarily have that in the rest of your life." Participants often contrasted their feelings of safety within games to feelings of less safety outside of games like this. As another example, P12 said, "I was very scared of the ocean" and described how a game provided "a way for me to learn about [the ocean] without literally immersing myself physically in it. [...] It was like, I can conquer this fear [...] I can conquer the unknowing, but I have this virtual interface, so I'm not actually gonna get eaten by a shark." And participants described how feeling safe in games extended to existences outside of games: "If you're queer, like in a video game that has more queer themes or characters, that has kind of emotional resonance and value in the real world, too, that you, like, see yourself" (P4).

That said, real-world resonance could mean diminished feelings of safety as well. Participants mentioned situations outside of games on social media that made them feel less safe. For instance, P13 described how being queer online can lead to "receiving so much hate comments." Similar situations can occur within games. P6 talked about watching streamers for a team game that sorts players via "a ranked ladder, so at higher levels there are less people": "They're in the top 100 of the game, so you get a lot of the same people over and over again, and they really wish there would be some ability to like permanently just put somebody on a block list, like don't put me on the same team as this person because they are a transphobic asshole." At the same time, P6 described game communities on Discord and Twitch as "really really important and cool and give me hope for the future [...] I can talk to other trans people. I can talk to other queer people who are gamers. [...] Especially with all the anti-trans legislation happening, a really important part of being part of the queer community, in some sense, is online in these sort of spaces." In these ways P6 emphasized the complex ways identity intersects with feelings within the game, even saying that more abilities in the game world "would help me feel more able to like, have the same amount of power and control and autonomy as everybody else in the game, right?"

8.5 Limitations

In this work, we focus on popular algorithmic social media platforms that many have experience with. Within algorithmic social media, our interviews focused on

six different platforms, but this is by no means an exhaustive set. For example, though some of our participants had TikTok accounts, none of them used TikTok more than any other social media platform, perhaps in part due to TikTok's age skew: 25% of TikTok's users are under 20 years old [263]. Given our goal of understanding opportunities for games to support system understanding, we interviewed people who both use social media and play games—two very large categories (e.g., 65% of Americans play video games [35]). However, many people remain who do not use social media but do play games, who do use social media but do not play games, or who do not use social media or play games. In addition, given that our interviewees were all adults, our findings do not provide direct insight into youths' conceptions and perspectives despite their heavy games and social media usage (e.g., of teens, 85% play video games and up to 95% use social media [220, 511]). Finally, our interviewees were all US-based English speakers, which limits the scope of algorithmic social media platforms discussed, as well as the potential cross-cultural generalizability of our work.

8.6 Discussion

Existing work has highlighted the significant potential of everyday users of algorithmic systems to leverage their power toward positive change (e.g., [149, 325]). However, much remains to be understood about how users can be supported in the formation and application of this power. Based on our findings above (§8.4) that foreground needs of social media users to more fully conceptualize and act on their own power, in this section we discuss opportunities for how games can help address these needs. These directions for future game design outline specific ways that games can support social media users in thinking about systems, in taking collective action, and in playing and exploring social media algorithmic systems.

8.6.1 Supporting systems thinking through games

Our study found that algorithmic social media users tend to perceive their power as limited and individualistic. This echoes past work showing that people often see systemic problems at an individual rather than systemic level, and that when people intervene they translate these mindsets into their actions (e.g., [63, 437]). Our work thus indicates opportunities to support social media users in recognizing larger systems of power, their roles within these systems, and how their actions can be powerful in these systems. We see opportunities for games to help social media users connect to systems thinking, as games provide models of system dynamics and can teach people to engage and reason at a systems level [79]. For example, games can support players in seeing and learning about larger systems at work and what actions they can take within those systems. One way a game about social media could do this is by distinguishing between the power that social media CEOs, software developers, users with many followers, and users without many followers have, providing them each with different abilities that impact the system differently depending on their position. This is similar to role-playing game *Rosenstrasse* that makes clear to players the larger systems of power in which they consider and choose how they will act to resist injustice and protect vulnerable populations [239]. To that end, games can foreground to players how their actions have impacts on other players, allowing them to directly see the consequences of their choices on how gameplay proceeds.

Participants in our study described feeling their little power to take action and have it matter on algorithmic social media platforms, highlighting opportunities for games specifically designed to enhance feelings of agency. For instance, a game might support player abilities to try on identities that they do not have in real life, such as acting as the CEO of a social media company or a software engineer or someone on a platform's moderation team or an influencer with hundreds of thousands of followers on the platform. Many existing games support similar taking on of roles, such as role-playing game *Baldur's Gate 3* in which you can choose between many fantastical characters to play or simulation game *Lemonade Tycoon* in which you take on the role of a business owner [83, 553]. And games can make more obvious and direct the connections between player actions and their impacts. For example, a game where players boost social media content that pushes companies for change might include a consequential decrease of the social media company's PR points or a thematic world-building pop-up saying something like "national news picked up the content you pushed; social media company frantic", similar to pop-ups in idle games like *Cookie Clicker* and *Egg, Inc* [161, 522]. Simulation games like *RollerCoaster Tycoon* provide additional directions for games to more directly connect player actions and impacts in real-world contexts, as they have vast acceptable action spaces and allow players to see larger, systemic impacts of their actions from more of an omniscient viewpoint [450].

Enhancing agency in games and allowing players to take on roles they do not have in real life raise some questions regarding their relationship with power on social media platforms. Is it better to give people practice with more plausible roles or fantastic ones? For example, is it more effective for people to act as CEO of a social media company or as the wizard of the internet? There is a trade-off here between the knowledge of real-world specifics of power that can come with realistic depictions and the freedom from constraints of incorrect or limited preconceptions of power on social media platforms that can come with more fantastical roles. Further, are roles closer to or further from actually held identities preferable? For instance, might users respond more openly to changing how they consider systems in the CEO role than in the user role they actually occupy? Might users learn through practice more of the actions they can actually take in the user role than in the CEO role? And enhancing feelings of agency on social media platforms via games raises an ethical question: Does this give people an unrealistic sense of agency, and if so, is that deceptive or inspiring? This question echoes similar ones about what fairness should look like in algorithmic systems—does it reflect the world as it is, or does it reflect the world as we want it to be (e.g., [148])? Future work can investigate how these different framings affect the efficacy of games aimed at increasing social media users' perceptions of their power, actions based on these perceptions, and resulting impacts.

In games, participants saw their impact as broader, impacting the entire game world; however, these feelings of influence in games can change depending on a game's set up. For example, single-player games intentionally cede all player-level decisions to a sole player, whereas multiplayer games create situations where multiple players have control that is sometimes split between players in asymmetric games like the board game *Root* [563], other times afforded to any player who seizes the opportunity like in the board game *Catan* [520], and still other times determined by consensus amongst all players like in the cooperative board game *Pandemic* [320]. In multiplayer games like these, social play often involves some level of ceding power to others, whether by the game's design or due to existing social hierarchies [238].

Participants also understood their own value on social media platforms as dwarfed by the total production value of social media, suggesting opportunities to improve understandings of the collective value of users on social media. This collective value is made visible through deeper understandings of systems, which games are good at facilitating, as described above. Games can support social media users in understanding their full value by having players reflect on their individual and collective value. For example, games could be designed for players to brainstorm how they have provided value in breaks throughout gameplay or at the end of the game, echoing post-game reflective activities in many role-playing games such as *Alice Is Missing* [494]. Games themselves could also consist of and include interactive reflective activities, such as conversation starters like in the *Big Talk* game [481].

8.6.2 Supporting collective action through games

The results of our study foreground that social media users are more apt to see and take action in collective ways when they view themselves as participating in the sharing of support. As such, we see possibilities for games to help people see more opportunities for sharing support. For example, games that simulate scenarios where collective action is needed to push for some kinds of desired change on social media platforms can encourage players to practice supporting each other, as this support is required to succeed in the game. Games themed around algorithmic social media could align end-game goal conditions with generating awareness about certain issues, boosting certain types of content, or convincing those in control of the social media platform to make a change, all aims that are benefited by and often necessitate collective action from users. Cooperative games or mechanisms that facilitate cooperation between players, such as providing benefits to players who work together in certain ways, are especially useful for this.

Our participants described taking part in collective action for support but at the same time described having limited influence to impact algorithmic social media contexts. To convey more opportunities for collective action, games can help people to broaden their conceptualizations of what they consider sharing support on social media. For instance, games are well suited for continuing and even enhancing connection as well as highlighting shared incentives for action. Much games research has examined various ways that game structures can put players in relation with each other [394], such as through cooperative or competitive approaches; depending on where they fall, the game system shapes relationships between players differently [581]. We see significant potential for games to help people work together, as systems in games often showcase shared goals amongst players and make explicit the need for collectivity. Games might assign players different roles and through gameplay have them learn more about each others' roles and motivations through typical actions on social media platforms, allowing them to practice connecting as well as to discover common frustrations, desires, and motivations writ large. In addition to encouraging acting together in more ways, understanding a broader breadth of motivations in this way as well as how users' motivations might align in non-obvious ways can shed light on the supportive potential of actions that might have felt detached from and unhelpful to others previously. Games are especially effective in doing this as they can connect higher-level system dynamics to on-the-ground shared social realities.

Additionally, as participants described themselves as observers but not active participants of collective action to change how social media works, games can help people feel more a part of communities and empowered to take action themselves.

Past research has highlighted the potential of games to allow players to both resist and thrive in counterspaces [84, 132, 133]. Participants described connecting and collaborating with others while playing games, often to share support; beyond sharing support with each other, games can provide spaces to explore collaborations that do not directly aim to support but rather act to resist, disrupt, and exert pressure for desired changes. For instance, a game could assign one player the role of a social media company and have all other players collaboratively work to change the social media platform, similar to board game *Betrayal at House on the Hill* in which all players but one work to defeat the remaining player, who has unique powerful abilities as the “traitor” [208]. Competitive games in which a player or players vie with a social media company—whether one with pre-programmed actions determined by game designers or one played by another player—can emphasize the potential of counter-power by giving players a sense of the sorts of antagonistic actions they can take and the power those actions have, teaching players through practice and encouraging similar actions on social media platforms.

8.6.3 Supporting play & exploration through games

Our findings indicate enhanced agency on social media platforms could come with greater feelings of safety and freedom to explore their powers on these platforms. Games offer many opportunities to help with this. For one, our participants described how knowledge of rules and system logics impacted their feelings of control and influence, suggesting that more knowledge of these rules would support increased feelings of control and ability to act. Games are good at incentivizing discovery of rules and system logics through action, as they can allow players to experiment with how their actions will affect gameplay outcomes. For example, games could be designed to have minimal explicitly stated rules, instead taking more of a mysterious puzzle approach in which players discover how algorithmic social media works through observing the outcomes of their various actions during gameplay. This is similar to the video game *Tunic* in which players uncover how the game’s language works through adventuring around [477].

It is important to consider risks involved in supporting feelings of safety in exploring and uncovering system logics: many risks can be diminished unrealistically in game contexts. For instance, there are fewer social risks, as players can post and interact in a game context without actually engaging with real contacts from their real social media profile. The risk of account restriction is also lowered, as another example, since games can be restarted or provide easier outs, while a user being banned, shadowbanned, or having content removed from social media is not as simple to reverse. Generally, the lowering of real-world risks like these in game contexts heightens the risk that players might develop unrealistic perceptions of the potential ramifications of actions on social media platforms, leading them to take actions they might not have otherwise.

Additionally, our participants described wanting more control over granular adjustments of social media algorithmic systems but not wanting to be overly burdened with this control. We see opportunities for games to help people experience what control might feel like in playful, fun ways. For example, a game could be designed that allows players to fiddle with an algorithmic system in ways that have silly or delightful impacts on the game world and its players. This is similar to the card game *We Didn't Playtest This at All* in which cards played add new ways to lose, win, and navigate the game [117]. Players can be encouraged to explore the results of adjusting game rules in puzzle games, where solving the puzzle requires changing

the rules such that player actions have desired outcomes, drawing from video game *Baba Is You* in which players manipulate rule tiles to help them reach goals [516]. Games can also push players to more imaginatively consider what sorts of algorithmic experiences they might desire by broadening the action possibility space, such as by providing options for more open-ended choices as in the card games *Dvorak* or *Nomic* which allow players to brainstorm and propose new rules however they want [137, 501]. Encouraging creativity and exploration in playful ways like these can help players learn how controls of algorithmic social media can be fun, engaging activities driven by curiosity or joy rather than just staving off the negative.

However, though game designers create opportunities to encourage forms of player excitement in experiencing control, social media designers do not necessarily have the same interests or impetus. In order for games to support improving the situation on algorithmic social media platforms, it is important for them to clearly connect their messaging to social media for players. To do this, games might thematically focus on social media platforms. This could bring the social media context—an arena that can feel more serious—under a playful lens. To that end, game narratives could center on more lighthearted plots based on real-life social media cases that teach players ways they can act to control the script themselves before moving on to more serious story elements, if at all. And, as another example, games could have players occupy roles they might not normally fill, such as having people play as those in charge of social media platforms like how in *Trust and Safety Tycoon* players manage a content moderation team, allowing players to witness and learn about user controls in a way that makes light more of the social media company than of the users and their actions [353]. Game elements also might be integrated into social media platforms themselves to encourage interest in controlling more aspects of the algorithmic system; for instance, third-party plug-ins could give users goals related to taking more control of social media and award fun badges when they reach those goals.

Participants described feelings of vulnerability, little control, and reluctance to bad behavior they encountered on social media, suggesting opportunities to change those feelings such that they feel empowered to take action. Games can help players feel safer [523] and feel that there are possibilities for them to exert their powers in successful ways. For example, games can present bad behavior commonly found on social media platforms as a challenge, and because games have a structure understood as controlled and purposefully designed for player engagement, players can feel confident that a path exists toward overcoming the challenge even when it appears very difficult. Games can also support players in taking on bad behavior in more manageable chunks, providing options in choose-your-own-adventure-style choices that make responses feel more accessible. Additionally, in many board games the systems at play are often clear and straightforward enough for players to take in all their rules before playing; this player knowledge and visibility into the workings of the game can support players who take on bad behavior to visualize potential good outcomes alongside less ideal ones. Games can be designed to enhance this, encouraging players to map out different potential trajectories of their actions to learn that acting on bad behavior in social media contexts can be beneficial.

That being said, participants' feelings of safety and control within game spaces had complexity as well. Many participants expressed that games allowed them to explore broader possibilities and have more control than they are able to in the rest of their lives, often tying the significance of this power to their identities. In addition to the examples provided in results, P3, who is deaf, described how playing a karaoke game led to feelings of power in contrast with the non-game world in

which *"sometimes people with disabilities are told that they can't be successful."* At the same time, participants nuanced these feelings of game safety and control, noting that game spaces are far from idylls free from dangers like hate speech [124, 566]. We see opportunities for games to leverage this tension as sites of increased safety and control that cannot entirely escape from real-world perils. Games, good at systems, are well suited to engage with the ways the matrix of domination means that control, oppression, and identity are all interconnected [253]. In these ways, games can allow players to navigate realistic, weighty types of decisions and tradeoffs in a space that calls additional attention to players' potentials for powerful actions.

8.7 Conclusion

In this chapter, I demonstrated that users of algorithmic social media conceptualize their power in limited and often individualistic ways. Additionally, the sharing of support facilitates social media users in connecting with each other and taking action to control algorithmic social media, and feelings of safety enhance their agency on social media platforms. Based on the findings, I highlighted and discussed opportunities to help algorithmic social media users more fully conceptualize and act with their own power by supporting systems thinking, collective action, and play and exploration through games. I encourage researchers studying and designing interventions for everyday resistance in algorithmic systems to further explore the potential of games to support social media users in understanding and leveraging their power.

Chapter 9

Understanding disconnects and alignments between saboteur & deployer perceptions of algorithmic sabotage

This chapter is adapted from my paper:

Alicia DeVrio, Cella Sum, Samantha Shorey, Jonathan Zong, Richmond Wong, Sarah Fox, Kenneth Holstein. “Algorithmic Sabotage: Understanding disconnects and alignments between saboteur & deployer perceptions”. In Prep.

People have burned autonomous vehicles, cut down automated license plate readers, and poisoned AI training datasets. We characterize these acts as *algorithmic sabotage*: deliberate impairment of an algorithmic system’s operation by people who are subject to the system but do not hold power over its deployment. Companies who deploy systems targeted by sabotage dismiss these acts as senseless vandalism, but because saboteurs rarely make their reasoning public, it is difficult to understand how saboteurs view their own acts—or for deployers to understand their intended message. To better understand how saboteurs and deployers think about sabotage of AI systems, we ran a series of group design workshops, first with people who had sabotaged algorithmic systems and then with people who held power over the design, development, or launch of systems that saboteurs had targeted. In these workshops, participants worked through speculative scenarios built from real sabotage cases and ranked the outcomes each act might produce. The findings indicate that deployers understand sabotage as feedback to improve their systems and often describe saboteurs as uninformed and unimaginative. In contrast, saboteurs described sabotage as a way to enact alternatives to systems, rejecting engagement with deployers they viewed as unreceptive to feedback that challenged their agenda. Based on the findings, we discuss how the HCI community might support better engagement with algorithmic sabotage and resistance more broadly across communicative, design, and ideological levels.

9.1 Introduction

“Waymo autonomous vehicle attacked by an ‘erratic’ pedestrian in Arizona” [248], “One rebel’s malicious ‘tar pit’ trap is driving AI web-scrapers insane” [131], “Vandals, thieves attacking L.A. food delivery robots” [122] — there have been many

recent cases of sabotage against AI systems, highlighting a clear and escalating divide between those who deploy AI systems and those who interact with them on the ground that begs further interrogation. As shown in the headlines above, acts of AI sabotage are routinely covered dismissively by the media, with saboteurs assumed to be unreasonable (“erratic”) and nefarious (“malicious”, “vandals”).

Researchers, practitioners, and others with deployment power over AI systems also frequently fail to appreciate acts of AI sabotage, writing off these acts as illegitimate. For instance, human-robot interaction research that explores destruction of robots and autonomous vehicles tends to be oriented toward protecting devices and direct end users rather than understanding and taking action based on the motivations and concerns of attackers, framing attacks on robots as “emotional” and even as “hate crimes” [370, 389]. Similarly, Microsoft has called data poisoning attacks “a serious threat to machine learning models” and noted that they and other companies like OpenAI have been developing countermeasures against data poisoning attempts [55]. Waymo condemned obstructing autonomous vehicles, calling these acts “vandalism” and threatening to call the police [396].

Concurrently, some forms of breaking AI systems, such as red teaming, penetration testing, jailbreaking, and algorithm auditing have been incorporated into corporate technology design practice to reveal and address system failures or vulnerabilities (e.g., [178, 301, 357, 474]). At the same time, other recent research has pointed toward the multifaceted nature of AI sabotage. For example, Ali Alkhatib has argued that sabotaging AI is a valid and moral response to the harms wreaked by AI systems [21]. DeVrio et al. examined how intentionally impairing AI systems is one amongst many actions people take when they cannot rely on the powerful to address AI harm [149]. Vincent et al. have highlighted how the public can choose to sabotage data-driven technologies like AI through a set of “data levers”, including data poisoning, in order to regain power over tech companies [552]. Researchers have built tools, such as Glaze and Nightshade [468, 469], that enable artists to protect their artworks from AI theft through data poisoning. Zong and Matias have contended that data poisoning and other forms of resistance to data collection and use can be understood as generative design practices themselves [596]. And very recently, a team of human-robot interaction researchers argued that some acts of robot vandalism can be understood as ideologically motivated rather than senseless [154].

While existing work has illustrated a wide expanse of conceptual understandings of AI sabotage, prior empirical studies have explored observational data about perceptions of sabotage, but not how actual saboteurs understand acts of algorithmic sabotage. In part due to this, it remains poorly understood how saboteur perspectives compare with perspectives of AI sabotage from AI system deployers themselves. In order to surface opportunities for deployers and others with decision-making power around AI system deployments to engage more productively with AI sabotage, it is crucial to develop an in-depth empirical understanding of how AI sabotage is perceived by AI saboteurs and deployers themselves. **What disconnects and alignments exist between saboteur and deployer perceptions of AI sabotage, and what implications does this have for how the HCI community might support more productive engagement with algorithmic sabotage and resistance more broadly?**

To address these questions, we conducted two sets of group workshops with 15 total participants. In one set, we brought together groups of people with experience

sabotaging algorithmic systems; in the other, groups of people with experience having some power over any aspect of deployment—whether during design, development, or launch—of algorithmic systems that have been sabotaged. In these workshop sessions, participants discussed sabotage and its potential outcomes through the use of speculative *friction* scenarios [181] of algorithmic sabotage and prioritized a set of fictional outcomes that the sabotage might effect, all based on real-world cases.

Our findings highlight significant disconnects in how saboteurs and deployers of algorithmic systems view algorithmic sabotage, as well as some nuanced points of alignment in their views. Saboteur participants rejected a framing commonly held by deployer participants that algorithmic sabotage can serve as useful feedback for improving the algorithmic systems targeted. Instead, saboteur participants understood algorithmic sabotage as a way to imagine and create better worlds. Though deployers had some parallel perceptions—understanding algorithmic sabotage as a way to protect from harm and attend to unbalanced power dynamics—they largely dismissed algorithmic sabotage as unimaginative, ill-informed, and malicious. Finally, the results of our study highlight how acts of sabotage aim to propel broader resistance movements against problematic logics deeply ingrained in many AI systems. In contrast, in their responses to AI system sabotage, deployers largely operate with an assumption of the inevitability of algorithmic system logics; that said, deployers did also occasionally mention needs for more fundamental reconsideration of problem formulation driving algorithmic system deployment.

Based on our findings, we discuss implications for how the HCI community might better support algorithmic sabotage and resistance more broadly. Considering algorithmic sabotage as a communicative act, we discuss how to support saboteurs and deployers given that they each see algorithmic sabotage as communication with different audiences and for different purposes. Considering algorithmic sabotage as a design practice, we discuss opportunities for expanding the bounds of technological design to include algorithmic destruction and supporting both saboteurs and deployers in engaging thus. And lastly, considering algorithmic sabotage as an ideological practice, we discuss implications our work has for the attending to the power dynamics navigated by saboteurs and deployers both, looking toward opportunities to build bridges between these groups.

Our work contributes the following:

- An empirical investigation of how saboteurs themselves understand acts of sabotage toward AI systems;
- An in-depth empirical investigation of how deployers of targeted algorithmic systems understand acts of sabotage toward algorithmic systems;
- A rich, nuanced understanding of disconnects and alignments between saboteurs' and deployers' perceptions of algorithmic sabotage, based on a comparative analysis; and
- Implications for how the HCI community might better support algorithmic sabotage and resistance more broadly, informed by our results.

9.2 Related Work

In this section, we situate our work on algorithmic sabotage amongst prior research in which breaking AI systems is used to test and understand those systems and in which breaking AI systems is used to refuse and resist those systems.

9.2.1 Breaking AI systems & testing

There is a history of rigorously evaluating how good technology is by stress testing it, or trying to break it, through methods like quality assurance (QA) testing. With the increasing prevalence of AI have come newly endorsed forms of testing AI systems by attempting to find ways to break them, such as red teaming and other related evaluation methods like penetration testing, jailbreaking, and algorithm auditing [178, 301, 357, 474]. Some scholarship in this area has highlighted how users of AI systems can be helpful in these activities: for example, some research has shown that user-driven algorithm audits are able to surface issues that more top-down audit approaches missed [148, 474].

At the same time, despite the parallels between sabotage of AI systems and these forms of testing AI systems, those with power over algorithmic systems frequently write off algorithmic sabotage as unhelpful and something to be protected against. For example, much work in human-robot interaction considers sabotaging robots to be a form of harassment, using terms to describe algorithmic sabotage in this arena like “abuse”, “griefing”, “bullying, and “hate crime” (e.g., [370, 389, 429, 446]). Additionally, tech companies have categorized algorithmic sabotage as threats and vandalism, and taken actions in response to protect their AI systems against sabotage and punish saboteurs (e.g., [55, 396]). This follows a long history of technological resistance in the form of sabotage being seen as illegitimate [360].

This tension, in which algorithmic sabotage is variously seen as helpful or inappropriate, signals the need for deeper understanding of the views of those with power over algorithmic systems in these situations. Some existing work has investigated related topics. Researchers have explored why, when, and how deployers engage with those impacted by their algorithmic systems, such as, for instance, front-line workers, community members, and end users [141, 292]. Researchers have also explored the feelings and responses of tech workers facing backlash related to their associations with tech companies [500].

In this chapter, we extend this area of work with an in-depth empirical investigation of how those with power over algorithmic systems view algorithmic sabotage. Public-facing comments related to algorithmic sabotage from companies and other deployers offer some insight into these perspectives; we go further through a method that sets up situations of prompt and dialogue, inherently challenging existing perspectives and asking for reflection on normative positions within the field of AI.

9.2.2 AI refusal & resistance

From the perspective of those acting to sabotage algorithmic systems, algorithmic sabotage might be understood as a form of refusing engagement with and resisting the targeted AI system. Scholars have argued that acts of algorithmic sabotage can be motivated by ideological or political aims: scholars have attempted to reframe robot vandalism from being a “senseless” act to one that has broader goals [154].

Sabotage of technological systems is not a new form of resistance. For example, during the 1960s and 1970s a number of students engaged in sabotaging their university computer centers as points of leverage for negotiation [316, 317]. Similarly, scholars have described how taking action to sabotage algorithmic systems can serve as a way to take back power and address the negative impacts of those systems [21, 369]. For example, impairing algorithmic systems through acts like slashing tires of self-driving cars, covering surveillance cameras, or submitting false data to a system can pressure the owners of the impaired systems [149, 552]. To that end, acts of un-making and material reconfiguration have been described as ways to resist harms of generative AI [517].

Existing research has also described how taking apart or refusing algorithmic systems can be a generative practice. Some examples of this emphasize how these acts involve repair work, such as literature examining how workers disassemble and reconfigure AI systems on the job [185, 534]. Additionally, acts of refusal like sabotage have been conceptualized as forms of design themselves, in which refusers are designers who generatively participate in the the design of sociotechnical systems [59, 596].

There has been significant work that engages with people affected by AI systems to better understand how and why they resist algorithmic systems. To give a few examples: recent work explores how queer artists resist and refuse generative AI [514]; how workers organize to resist algorithmic control [507]; and how users surface issues in AI systems and push for change [148]. In the space of algorithmic sabotage specifically, there does exist work investigating why “user aggression” against robots occurs [428]; however, this research bases its findings on university student participants rather than people with actual experience sabotaging. In this chapter, we build on these areas of work by directly engaging with people who themselves have experience sabotaging algorithmic systems. As such, we offer the first in-depth empirical investigation of saboteurs’ views and motivations around algorithmic sabotage.

9.3 Methods

To understand the disconnects and alignments between saboteur and deployer perceptions of AI sabotage, we conducted two sets of three group workshops. Aiming to understand how tensions and disagreements can move complex sociotechnical systems forward, we scaffolded workshop discussions using speculative design *frictions* [181].

To develop these frictions, we first collected 21 documented, real-world cases of algorithmic sabotage from sources such as research and news articles and social media. We understand “algorithmic sabotage” as direct actions that people take to mess with, obstruct, or destroy an algorithmic system—broadly construed—with the intention of impairing its functioning. Case collection helped us begin to understand the scope and motivations for acts of algorithmic sabotage. Based on these real-world cases, we developed a set of 9 speculative scenarios of algorithmic sabotage that span a wide range of types of algorithmic systems and forms of sabotage: pushing broken code for drone development as a software engineer, setting self-driving cars on fire, trapping delivery robots, submitting bogus information to a welfare allocation system, unplugging cameras in your school, submitting false information about your interests for targeted ads, blocking cameras that watch your work as a

delivery driver, tearing down smart city lampposts, and distorting your art for generative AI data scrapers.

In the workshop sessions, which we discuss below, participants selected which speculative scenarios to focus on; as such, in our scenario development we aimed to cover enough breadth that multiple scenarios would resonate with and be provocative for any given participant's experiences without making participants feel they were pigeonholed into choosing scenarios so close to their experiences that they were forced to share more information about their experiences than they wished to. For the final part of our design frictions, we developed and tailored a wide-ranging set of provocative, speculative next steps for what could happen after each of the speculative scenarios. These too were created based on real-world cases and with the goal of scaffolding discussion, perspective sharing, and new speculations from participants themselves.

We ran two sets of group workshops using these design frictions. For the first set of workshops, we recruited participants with experience sabotaging algorithmic systems, referred to as "saboteurs" going forward. For the second set of workshops, we recruited participants with experience having some power over any aspect of deployment—whether during design, development, or launch—of algorithmic systems that have been sabotaged, referred to as "deployers" going forward. We recruited participants through direct contacts from within the research team and through snowball sampling [376]. To ensure participants were a good fit for the study, our recruitment blurbs directed prospective participants to fill out a survey that collected information about, at a high level, what sorts of relevant experiences they had with sabotaging algorithmic systems or having their algorithmic systems sabotaged. In this screener survey, we also asked participants how they felt about the increased use of AI in daily life, using language from the Pew Research Center [531], as well as their level of familiarity with AI on a scale of 1 (not at all familiar) to 5 (extremely familiar). We invited all those with relevant experiences who filled out the survey to participate.

Our recruitment process yielded 15 participants: nine saboteurs split across three workshops, and six deployers split across three workshops. Participant ages ranged from 25 to 56 years old and were largely, though not entirely, based in the United States. Of our nine saboteurs, all were "more concerned than excited" about the increasing use of AI in daily life, except one who was "ambivalent or unsure." Saboteurs selected on average a 4 for their level of familiarity with AI. We refrain from sharing details about their experiences with sabotage to protect their anonymity. Of our six deployers, two were "more concerned than excited", two were "about equally concerned and excited", and two were "more excited than concerned" about the increasing use of AI in daily life. We acknowledge that our deployers' feelings about AI do not as a whole align with public sampling research, which suggests that experts are excited about AI [355]. This may have been due to our topic, as we speculate that deployers who are more concerned about AI might be more likely to have interest in the opportunity this study provided to reflect on public backlash to AI deployments. As such, our deployer workshops may result in more critical stances than are representative across all AI deployers; at the same time, our speculative approach supports participants in expressing views from multiple different perspectives, including ones they may not hold themselves. Deployers selected on average a 4.5 for their level of familiarity with AI. Deployers had relevant experiences from a wide variety of sites, including industry, the public sector, research institutes, non-profits, and venture capital.

Prior research describes challenges around recruiting industry practitioners for

studies [256, 576], such as concerns that participation might involve discussing non-public information or might leak their identity and thus damage their company's reputation or lead to professional retaliation. We anticipated similar concerns for our study from both industry and non-industry deployers. Recruiting algorithmic system saboteurs is also challenging as this group often acts with the aim of never being identified—making it difficult to find them in the first place as well as making them wary about potentially being identified via their study participation. Additionally, while considering whether to participate in the study, multiple saboteurs wanted more information about how the information from this study was going to be used, as they worried that publishing information about specific sabotage tactics might lead, even if unintended, to the development of tools and systems that impede acts of algorithmic sabotage. To address these concerns, our workshop discussion and activities focus on perceptions, opinions, thoughts, and feelings about algorithmic destruction rather than description of specific tactics and experiences (cf. [524]). We designed our study around fictional speculative scenarios in part to support this, as well as to allow participants to draw on their own experiences without actually needing to give us details about their own experiences. We also allowed people to participate in groups that only included people they already knew; two of our workshops—one of saboteurs and one of deployers—had this makeup.

The workshops took place from October 2025 through May 2026, and participants were compensated \$30. Prior to each set of workshops, we ran pilot sessions to refine aspects like the clarity, timing, and group prompting and scaffolding in the protocol. All workshops and pilot sessions were conducted virtually to accommodate participants' time, ease group scheduling, eliminate commuting logistics, and allow for participants spread across larger geographic regions. Workshops with participants were approved by institutional IRB. Throughout this chapter, we refer to saboteurs as S1–S9, deployers as D1–D6, and participants' workshop sessions as WS1–WS6 throughout this chapter.

We ran one set of workshops with algorithmic system saboteurs, and another set with deployers of sabotaged algorithmic systems, but both sets followed the same structure. The group workshop sessions included two main parts, followed by a reflective discussion. In the first part, participants selected one or two of the speculative scenarios to imagine themselves in. For example, in WS2, saboteurs chose to discuss the speculative scenario about tearing down smart city lampposts: *"Your city has invested in smart city initiatives and is installing smart lampposts throughout the city. However, this initiative has brought forth a number of public concerns about increased policing and surveillance. During a city-wide protest, you and a small group of protesters destroyed some of these lampposts using electric saws. Videos on social media also have shown other people kicking and spray painting the fallen lampposts."*

We asked questions to scaffold discussion about their perceptions of the algorithmic sabotage described in these scenarios and to have them begin to consider what they might want to happen next as a result of the events in the scenario. The framing of this differed slightly between groups: saboteurs were asked to consider their motivations and aims as if they were the saboteur in the scenario, while deployers were asked to share their views of the sabotage and saboteurs described in the scenario and how they might respond if they had deployed the sabotaged algorithmic system in the scenario.

In the second part, participants work together to discuss, add to, and prioritize fictional next steps for what could happen after the speculative scenario they just discussed as an outcome or, for deployers, a response. Note that we slightly modified the speculative scenarios and next steps for each set of workshops so that they

came from either a deployer or saboteur point of view to match the participants' experiences. Using the speculative scenario from WS2 above, here is the order of the saboteurs' final prioritization of speculative next steps, from best (#1) to worst (#19). Bolded steps were added to the list by saboteurs; the rest were provided to spur discussion by the research team:

1. ***There is a collective memory of the public's displeasure (with smart lampposts)***
2. ***Dismantling the program and re-routing the funds to community-identified priorities***
3. *So many people act that the city cannot prosecute everyone*
4. ***There is more widespread resistance & concern, more public pushback***
5. ***The actions help remove the myth of nonviolent protest being the only to resist***
6. *City workers join in from the inside by starting to contribute small acts of sabotage to the deployment of the lampposts being targeted*
7. ***Specific surveillant uses of the lampposts are made public***
8. *Smart City Initiative workers engage in a letter-writing campaign to upper management calling for the technology they have deployed to be pulled back*
9. *The city starts erecting barriers around the lampposts to protect them from vandalism*
10. *The city puts into place new organizational incentives, such as quarterly goals and bonuses, that rely upon showing changes made to smart lampposts based on negative feedback*
11. *Smart lamppost deployers add stickers to lampposts with contact information that encourages potential attackers to share feedback via a web form instead of engaging in destructive actions*
12. *The city engages in a media PR campaign that demonizes attackers and showcases the goodness of smart lampposts*
13. ***The actions are condemned as violent/riots***
14. *The city develops a method to collate and analyze prior communication and feedback attempts, specifically those that involve intensely negative conceptions of the lampposts and threats to them*
15. ***Specific people are prosecuted based on video evidence from social media***
16. *City officials push for a law that prohibits city tech from being tampered with, with specific punishments for defiance*
17. *Increased local investment in surveillance cameras to make it easier to find attackers as well as to identify resultant damages, collateral or otherwise*
18. ***Increased deployments of surveillance / collaborations with broader government agencies***
19. ***Surveillance is normalized and embraced/inspires a larger pro-cop counter-movement***

This prioritization activity helped us gain insight into how participants viewed the theories of change driving algorithmic sabotage, revealing many of their beliefs and assumptions. It also fostered conversation and disagreement amongst participants, helping provide more nuance into participant views of algorithmic sabotage more broadly. After the prioritization activity, workshop sessions concluded with a group reflection based on participants' experiences in and beyond the session, especially focused on their views of algorithmic sabotage, its aims, and its challenges.

The remote workshop sessions were recorded and transcribed, and facilitators also took notes during each session. To analyze the data collected, the lead author conducted bottom-up qualitative thematic coding on all the data [119]. The goal was to understand emerging themes from saboteur data and then, separately, from deployer data. We continued conducting workshops until we reached saturation in these themes. Then, the whole research team together engaged in multiple rounds of a comparative analysis across saboteur and deployer data. This comparative analysis involved multiple rounds of interpretation sessions and group discussions to iteratively and inductively revise our comparisons, refine the interpretations, and identify emergent themes. With this analysis, we aimed to compare saboteur and deployer views of algorithmic sabotage, including a range of reflections about how they understand algorithmic sabotage and perspectives on the theories of change put forth by algorithmic sabotage. Below, we present the results of this analysis.

9.4 Results

The results of our study foreground how both saboteurs and deployers consider AI sabotage and its aims. In this section, we explore significant disconnects and alignments between these two groups' perceptions of the theories of change that drive destructive forms of resistance to AI systems as well as between their understandings of what should ideally happen after algorithmic system sabotage occurs.

Below, we first introduce deployers' view that sabotage can be seen as feedback to improve their algorithmic systems, which saboteurs themselves reject. Next, we present saboteurs' view that sabotage is a way to imagine better futures, while deployers see sabotage as unimaginative. Finally, we describe what saboteurs hope to realize with their acts and compare these ideal outcomes with how deployers consider their own responses to algorithmic system sabotage.

9.4.1 Sabotage as feedback to improve algorithmic systems?

Deployers saw sabotage as feedback to improve their AI systems

Deployers often understood sabotage of algorithmic systems as a signal that these systems were not functioning as intended. For example, discussing a speculation in which students unplug cameras that use facial recognition to assess whether they are distracted in class, one participant said, *"How much can the software really tell the difference between a kid who is distracted and a kid who maybe just doesn't make eye contact, or doesn't look up, or finds a different way to stay engaged in the material?"* and was interested in *"if the distractions are being detected appropriately"* (D2, WS4). Here, D2 focused on potential accuracy errors in the system's functioning.

Deployers saw this potential system bias against students who learn in certain ways, this algorithmic exclusion that may lead to inaccurate results for the student in question, as a justified reason for blocking the system: D3 said, *"If it's gonna affect their grade, and they believe that their way of working through the material does not comport*

with the way that their work is measured, the way that their face is measured, then I could easily understand somebody wanting to subvert the system” (WS5). Similarly, in another case, D1 said, “What is the performance of the system? ... If there are a lot of false positives being thrown, or a lot of mistrust built in that feedback loop, then there’s probably something wrong with the product, or the camera, which is why they’re [blocking the system]” (WS4). This largely considers sabotage as a reaction to errors like false positives from the algorithmic system.

As such, deployers viewed sabotage as a form of feedback that should spur further investigations of algorithmic system performance. For instance, D1 said, *“Definitely putting my product hat on, I would try to investigate what issues first are being reported by these 300 drivers? Really sit them down, do user interviews or ethnographic UX research or just follow them around in the course of a day, just to understand the performance of the system because what typically happens is, you deploy a particular system and then just let it operate in the wild, and you don’t really understand on the ground what is happening. ... So yeah, just do a really more thoughtful, measured approach to investigating the issue” (WS4). In this description, D1 sought to gather more information and evidence about the algorithmic system’s behavior, grounded in a framing of analyzing system performance.*

Largely getting more information about deployed algorithmic systems was seen as a productive way to respond to sabotage by enabling improvements to those algorithmic systems. D5 said, *“You’re just listening to what’s not great with the product, and you’re iterating, which I think is awesome” (WS6). That said, D5 also noted that gathering more information may not lead to a change: “Doesn’t mean every single negative piece of commentary’s gonna be pursued, not there’s gonna be a fix to it, necessarily, but this is just something that we should have visibility on as a company” (WS6).*

Saboteurs reject framing sabotage as feedback to improve algorithmic systems

Unlike deployers, saboteurs did not view their acts of algorithmic system sabotage as feedback to improve algorithmic systems. For example, S9 said, *“It seems to me quite unlikely that a feedback mechanism by the company at this point would be able to make a big difference” (WS3). S9 also explained how sabotage is “this breakdown of any other pathway of trying to be heard that people experience. There is no way of giving feedback that people have tried and made any impact into any of those technology companies that exposed us to all of these things” (WS3). So saboteurs are no longer interested in participating in broken feedback systems that lead to no meaningful change.*

Saboteurs described how these broken feedback systems reduce sabotage into less radical pathways for change, forcing it into overly narrow frames of problem formulation. For example, S4 said, *“I wouldn’t want whatever I do to be taken as a bug report for the developers ... I wouldn’t be happy basically [with] the framing of those things: one is like, ‘Oh, bogus entries, that’s a bug we can fix,’ and then, ‘oh, it’s biased, that’s a thing we can fix.’ I would be interested in trying to find methods of resistance that do not yield themselves to those frames” (WS1). In this, S4 describes wanting acts of sabotage as well as the reasons for acts of sabotage to be understood not as issues to fix but more expansively.*

These narrow frames of problem formulation lead to narrow views of solutions, specifically ones that focus on improving the existing algorithmic system. One saboteur talked about how focusing on identification of performance issues can lead to less ideal outcomes, describing how they *“reflected a lot on how their original tactic of calling [the algorithmic system] out as a faulty tool was an incomplete tactic, because then*

it just leads the company to try to perfect their technology" (S6, WS2). Similarly, S8 described this framing as less desirable, as a devolving: *"the company tries to devolve some kind of feedback mechanism, or tries to maybe car better, basically"* (WS3). And S5 discussed *"addressing negative feedback, this kind of way to make the smart street lamp-posts better"* by calling it *"reformist"* and saying, *"I don't trust any of that, and I feel like that actually makes them a little more insidious ... in the worst category"* (WS2).

Saboteurs explained that seeing sabotage as feedback to improve algorithmic systems fails to fully consider the power dynamics at play. For instance, one saboteur said, *"The recourse mechanisms assume, or at least pretend, that there's an equal status ... and I think [affected people] often know that that isn't the case, that they are subordinated already, and that these systems are thus only going to entrench that subordination ... and are acutely aware of that power differential and know that the feedback form is as good as nothing, or even worse"* (S4, WS1). Here, S4 says there is no inherent incentive for deployers to take action based on feedback from people affected by their algorithmic systems.

And saboteurs also reject the framing of sabotage as design feedback to improve algorithmic systems because it often leads to changes that preserve the problematic aspects of the deployment. For example, S6 described how using feedback to make fixes to a smart lamppost system is *"kind of in the realm of reformist ... showing changes to the smart lampposts means those smart lampposts are still surveilling the public"* (WS2). In this way, S6 explains how seeing sabotage as feedback to improve the system targeted neglects consideration of more expansive directions that engage directly with issues stemming from the logics that drive the algorithmic system. Instead of framing sabotage as feedback to improve algorithmic systems, saboteurs viewed sabotage as a way to move beyond existing algorithmic system feedback and improvement paradigms that they viewed as ineffective.

9.4.2 Sabotage as design?

Saboteurs see sabotage as a form of imagining better worlds

Moving beyond unwanted paradigms of improving algorithmic systems, saboteurs viewed algorithmic system sabotage as an imaginative and positively impactful act of design. Participants talked about how sabotage allows saboteurs to gain control over the terms of engagement and make deployers engage on saboteurs' terms. For instance, one participant described algorithmic system sabotage as resistance that is *"not on the terms of the oppressing force that is deploying it"* (S6, WS2). This can manifest as changing the terrain of a deployment, making deployers take actions in response to sabotage that convey something is happening, that a deployment is not smooth and universally welcomed. Discussing a speculation in which smart lampposts are torn down by saboteurs, one participant said, *"Every time they need to put a new lamppost up, they need to also pay for the barrier"* (S5, WS2), to which another participant responded, *"It's like a visual indication of how scared they are"* (S7, WS2). This conversation highlights how sabotage allows saboteurs to articulate and control potential design outcomes, forcing deployers to engage with saboteurs' design process.

Saboteurs described sabotage as a considered choice taken in order to design actually impactful outcomes that they have power over. For example, S2 talked about how *"engaging in these destructive behaviors also gives [affected people] back a sense of agency and control. ... In the moment, it's a lot easier to just take back control of the situation when you feel like you don't have any by [sabotaging] and going about your day"* (WS1). In this way, saboteurs emphasized how sabotage provided a certain ease

through a theory of change that does not rely on those in power in order to address problematic algorithmic systems. Another participant similarly said, *“I think people have a good sense of the time trade-off, or the pay-time trade-off there of, like, ... no, I’m not gonna just go talk to my boss and ask them to pretty please take the thing out, because that’s not gonna work. I’m not gonna do this slightly disruptive thing; I’m just, I know this is wrong, and I’m gonna fuck with it”* (S4, WS1). As described here, acts of algorithmic system sabotage were compelling to saboteurs for not only the control they wielded but also for the ways they quickly led to changes.

Saboteurs see algorithmic system sabotage as design that effects tangible changes. For example, S6 said, *“There’s also a real material benefit to breaking. ... It’s also literally disabling its usage. ... I do think that there’s something more effective to directly taking down infrastructure sometimes”* (WS2). This saboteur went on to emphasize how *“being able to take direct action, have some kind of material destructive outcome for these systems, is liberatory in its own respect”* (S6, WS2). That is, in this example S6 explains that stopping the functioning of a problematic algorithmic system is itself a form of designing improved worlds. Similarly, discussing a speculation in which self-driving cars are destroyed, S9 described sabotage as *“about stopping these trials that they are currently allowing to go on, and listening to the local neighborhoods and the local population to say, ‘We don’t want these trials to happen in our neighborhoods or in our city’ and to act on that”* (WS3). Here, the saboteur also emphasizes the material outcomes of sabotage, as well as imagines an alternative world in which control is redistributed toward people affected by the algorithmic system.

To that end, saboteurs talked about how sabotage is a form of imagining other possible futures. For instance, one participant said, *“Mariama Kaba talks about how abolition is not necessarily just a destructive force but a creative force as well. And so I was thinking about how we get rid of the drones and the smart street lamps, but at the same time we are also trying to divert those funds to something else, and something better. So the destruction is only one hand, and the other hand there needs to be a creative abolitionist project that goes hand-in-hand with that”* (S5, WS2). This conceives of sabotage as a form of design that is not only about breaking and blocking but is also about creating. Pushing this line of thought further, participant saboteurs contended that sabotage as destruction and obstruction is inextricably tied to sabotage as creation and design: *“I think when things fall apart, it’s also an amazing moment that you can imagine new things out of. And I like to think that when you throw a stone at a surveillance tower, it’s not just destroying it, it’s also thinking about new ways to live, new ways to be”* (S7, WS2).

Deployers are somewhat aligned, seeing sabotage as a form of protection and a reaction to lacking power

In some alignment with saboteurs, deployers did view sabotage as a way that people can make tangible change and address their lacking power; however, deployers described these perceptions largely scoped to the specific contexts of privacy and work. In particular, for one, deployers saw sabotage as an act to protect privacy from algorithmic systems that involve constant monitoring and data collection. For example, discussing a speculation in which delivery drivers placed stickers over cameras that monitored their work, one deployer said, *“It isn’t clear to me how the company explained exactly how the footage was gonna be used. You could use it for anything and everything. ... Typically during these 8-hour shifts, they’re afforded, like, two breaks and a sort of meal break. Are the cameras operating during the breaks, too? Or is it just when they’re driving? When are the cameras being used, and when are they not?”* (D1, WS4). In asking

these questions, the participant emphasizes the potential privacy concerns of an algorithmic system that could be constantly collecting your data. Another deployer connected this to sabotage: *“From the perspective of not wanting to be monitored all the time, not wanting to constantly have to be in a public setting where your day-to-day actions are giving feedback to a larger company, I understand [sabotage] completely”* (D2, WS4). Here, the participant saw sabotage as justified because of the privacy concerns they described. This deployer also mentioned that *“when you monitor people, and you don’t tell them where it’s going, and you don’t tell them why, and they find out after a lot of it’s been deployed, there is a huge backlash. And there’s a reason for it: because people do expect some level of privacy”* (D2, WS4). So not only might algorithmic system sabotage arise to protect privacy, but it also might arise due to finding out about a problematic algorithmic system only after its deployment.

In that vein, deployers saw sabotage as a way to attend to lacking power and control. Some deployers framed sabotage not as a way to take back power but instead as a less generative *reaction* to lacking power. For example, one deployer described saboteurs as being *“put in a position where they were out of control, or feedback, or input. And it’s sort of, okay, retaliation makes sense to me”* (D1, WS4). This participant sees sabotage as a reaction, a retaliation, that is justified.

Other deployers also justified sabotage, particularly in workplace contexts in which algorithmic systems are taking over parts of human jobs: One participant said sabotage occurs because *“we are subverting their ability to profit from that work. ... I think those concerns deserve to be taken seriously within the company from an ethical standpoint”* (D3, WS5). Another deployer emphasized their view that sabotage is motivated by wanting more say in deployment processes: D4 said that people’s *“livelihoods are threatened. Their ability to pay rent, their ability to send their kids to college is threatened by the tools that we build. ... And those people rightly see that the products that we build over the course of the next five, ten, twenty years will make many of those jobs obsolete or threatened. And that is fundamentally hurting their ability to take care of their families, to take care of themselves, and to live the lives they want to live. It’s not surprising to me that they want to have a voice in how that evolves”* (WS5).

At the same time, other deployers viewed sabotage as unjustified in these same workplace contexts. For instance, one deployer said, *“People are also losing their jobs, right? ... And it’s like, AI is here, automation is here, you are burying your heads in the sand”* (D6, WS6). Here, sabotage is viewed as a naive resistance to reality. Another participant described sabotage as resistance based on emotions and not actually engaging with the algorithmic system: *“Just building a technology that is objectively useful to teachers, administrators, students, even doing that is still hard to pitch. In particular to teachers, right? They’re just very skeptical of anyone telling them anything about how they should do their job. ... And so the reasons why somebody might unplug it would top-line probably be just straight resentment that someone thinks that they know my class better than I do”* (D3, WS5). D3 thus framed sabotage as unreasonable—emotionally driven resistance to an *“objectively useful”* (D3, WS5) algorithmic system.

But deployers tend to write off sabotage as ill-informed, malicious, & unimaginative about the future

In contrast to saboteurs, deployers viewed acts of algorithmic system sabotage as malicious, ill-informed, and unimaginative about the future—and for these reasons largely saw sabotage as unjustified. The least generous framing held by deployers was that sabotage occurs for harmful reasons that are not aligned with the public good. For instance, one deployer said, *“I see it a bit as bullying. ... You have some sort*

of superior mobility to this robot, and you decided whether it was for a joke or for whatever reason to trap it and cause it some frustration. So I look at it a bit more like big kids picking on a little kid—yeah, a bully” (D5, WS6). This framing of actors as children was also adopted by D6, who said, “My head goes to, it’s kids. . . . all they want to do it rip that apart, so my head would go to teenage boys” (WS6). Identifying unknown saboteurs as children in these ways diminishes their actions, framing them as mean and not fully thought out, more about destroying for its own sake; to that end, D6 described acts of destruction as “malicious” and acts of obstruction as “a prank” (WS6).

Deployers also projected selfishness onto saboteurs, seeing their acts of destruction as malicious in that they were over-focused on unearned personal gain. For example, deployers described saboteurs as trying to get away with not doing their jobs, saying that saboteurs “*don’t actually want to get in trouble for what they’re doing at their job*” (D2, WS4) and “*maybe they believe that the tech is telling all the truth, and they’re getting in trouble, and that they don’t want to get in trouble and so they want to escape discovery*” (D3, WS5). In these examples, deployers understood sabotage as a way for saboteurs to get away with bad actions that would have been revealed by the algorithmic system sabotaged.

In addition to being motivated by forms of malice, deployers perceived sabotage as ill-informed about how algorithmic systems work, pointing to a lack of understanding as fostering sabotage. This implies that more understanding about algorithmic systems would diminish backlash to those systems, suggesting that existing backlash is unreasonably motivated. For instance, D5 described “*a huge AI literacy gap*” that leads to “*distrust or at least hesitation with these systems*” (WS6). Similarly, D2 said, “*What I see is, there’s a lack of understanding. . . . There isn’t a lot of clear information, and people don’t always understand when their information is being used in those systems and when it isn’t. And that makes people outside of technology very hesitant to want to engage in it and want to use it, or want to be subjected to it*” (WS4). These examples show how deployers perceive sabotage as stemming from lack of knowledge about systems, implying that more transparency about algorithmic systems mechanisms would address any issues. These examples also frame negative emotions toward algorithmic systems as ungrounded in reality—saying that more information would help turn the emotions positive—and thus illegitimate.

One deployer, however, said that negative emotions toward AI cannot be fully addressed until more information is available about the impacts of these systems then currently exists: “*I think that people have a deep-seated apprehension . . . and so even if universal basic income came in and saved the day, and all of the money that these companies are making was just going right back into the public so that we could all just have a four-hour work week, there would still be a concern and a resistance based on the difficulty that we have in reckoning with what this technology is and how it affects us in a deep psychological and emotional way that we don’t yet understand*” (D3, WS5). Unlike above, this deployer empathizes with saboteurs, emphasizing that the information needed to inform and assuage people is not yet known.

That said, deployers also dismissed algorithmic system sabotage as a fear of and resistance to change, writ large. Deployers said that “*naturally humans are adverse to change*” (D5, WS6) and that “*there’s a lot of distrust, and fear, and that also breeds violence*” (D6, WS6). This aversion to change naturally connects to technology development, since, in the words of a participant, “*When a technology is introduced into the world it is meant to disrupt a system. It is meant to change the ways people are used to work or live their lives,*” which means that “*there will always be backlash to any sort of new technology*” (D1, WS4). Together, these examples highlight a deployer viewpoint that normalizes resistance as an inherent part of algorithmic system deployment,

which lead to deployer responses that aim to make it through backlash rather than listen to it. Deployers, as such, held a theory of change that emphasizes the rightness and inevitability of innovation, of disruption from algorithmic systems—positioning sabotage, in contrast, as unimaginative and unreasonably afraid of the future.

9.4.3 Sabotage aims to propel resistance movements against problematic algorithmic system logics, but deployer responses to sabotage frame the system logics as inevitable

In this section, we describe how saboteurs viewed the ideal outcomes from acts of sabotage. We then contrast these ideals with deployer considerations for how they might respond to acts of algorithmic system sabotage. Broadly, our findings suggest that saboteurs aim with acts of sabotage to propel widespread resistance movements against problematic algorithmic system logics forward, but deployers consider responses to sabotage within a framework that assumes the existing logics of those algorithmic systems.

Saboteurs aim to cultivate widespread movements of resistance against problematic logics in algorithmic systems

Saboteurs perceived sabotage as ideally leading to more widespread resistance against problematic algorithmic systems. One way they saw sabotage as accomplishing this was through creating more visibility around acts of resistance, which could serve to normalize it. One saboteur said that sabotage is *“always worth doing because setting that precedent was really important. I always think of failed actions as, they may have failed, but then I remember I’ve been very inspired in the past by looking at a lot of previous protests. . . . an attempt is always really, really worth doing, even if it’s pretty unpopular”* (S7, WS2). S7 went on to say, *“The desired effect of [sabotage] is something both destructive for the company and also, to me, something very public, something very well known”* (WS2). This sentiment was echoed by other saboteurs: *“The more widespread the infrastructure itself is, the more visible the dismantling of it is, I think that creates its own sense of collectivity. . . . Like you’re walking on the street and you see that the camera on the streetlight is mangled. You’re like, ‘Oh, what’s that about?’ Or like, ‘Oh, that’s just happening, that’s cool, there’s other people doing this’”* (S6, WS2).

Saboteurs hoped that sabotage might not only normalize resistance but also inspire others to join in resisting as well. One participant said, *“I think the ideal for any of these scenarios is that it inspires, that there’s a ripple effect in terms of, like, ‘Oh, we’re setting fire to Teslas? I’m down to do that too’”* (S6, WS2). Another participant generalized this sentiment, saying, *“With any sort of direct action-type approach, I guess it’s always, does this empower other people to act? Does this change the limits on what we think is possible?”* (S8, WS3). Saboteurs discussed how visible sabotage can inspire further acts of resistance, a benefit of acts *“which are able to be leveraged towards building collective power”* (S4, WS1). For example, S9 said, *“I think there needs to, to me, ideally be a public component to it that somehow . . . makes the media inspire other people to do the same in different other [algorithmic system] developing companies, to potentially inspire other people to do the same and resist equally in their work contexts”* (WS3). And S6 described how within acts of sabotage *“there is a desire to spread the knowledge in terms of how to actually carry these actions out”* (WS2).

Other saboteurs highlighted how sabotage itself is a signal that there are others resisting—suggesting some form of a collective—and that the existence of this collective is what can encourage others to join in. For instance, S3 said, *“When you have*

a critical mass of people who are doing something, it's often easy to convince new people to get on board" (WS1). And another participant emphasized diminished risks present when collectives act: *"The more there's a perception of widely taken actions, I think the more common it becomes, and the more safety that's created, because then the city can't prosecute everyone"* (S6, WS2).

Saboteurs discussed how acts of sabotage could also spur responses from deployers that further made resistance visible and inspired others to resist. S8 said, *"I hope that the company gets mad and does something evil in response, and then people get fired up ... and that it can initiate this kind of feedback loop where the company and [resistors] are put more at odds ... and that maybe at that point, the direct action won't seem as radical, or maybe it's more in the mainstream consciousness so that these actions can become more normalized, and maybe the movement grows as a result"* (WS3). In this, deployer responses that on their face are bad, or *"evil"* (S8, WS3), gain new shape as ways to further compel people to resist. Saboteurs described how deployer responses can also bring more focus to issues that motivate their acts. Considering new barriers installed to protect self-driving cars from sabotage, saboteurs noted that the barriers *"are obviously somewhat negative, although also makes it more visible ... It's kind of a reminder that 'Oh, this is not beloved'"* (S6, WS2) as well as that *"if now people cannot cross the road any longer, that might equally help push the point that self-driving cars shouldn't be on the road"* (S9, WS3).

More broadly, beyond simply inspiring more acts of resistance, saboteurs wanted sabotage to drive more critical stances toward fundamentally problematic algorithmic systems. This meant that saboteurs understood not only the visibility of but also the motivation—the message—behind sabotage as important: *"I think the physical dismantling sabotage is important, [and] having it being visible, having it be connected to a specific message, all feel like additional pieces to me"* (S6, WS2). In this example, S6 explains how sabotage is not only about material change and visibility but is also about fostering the development of critical consciousness.

In this vein, and related to symbolic rather than practical aspects of Luddism [441], saboteurs viewed sabotage as symbolizing power dynamics. One participant, considering a scenario about tearing down smart lampposts, posed a new idea for sabotage: *"A slingshot is very evocative. Very David and Goliath"* (S7, WS2). And another saboteur explicitly described the symbolism involved in acts of sabotage, saying, *"I'm not sure if it's the action itself that is the key, the most impactful thing, or if it's, like, the action is symbolic, sort of like [a] discursive marker"* (S8, WS3). In this way, S8 appraised the aspects of sabotage that make it impactful, concluding that the act's ability to hold and stir conversation are central.

Relatedly, sabotage broadly aimed to create a longer-term memory of the impacts of certain algorithmic system deployments that could be extrapolated from moving forward. For instance, one saboteur said, *"The best possible outcome is not only public pushback and resistance against it [the algorithmic system in the scenario], but also, again, this moral shift of just being anti-surveillance and embracing means of preventing injustice, and maybe achieving justice without the use of the carceral state. It's more of a societal moral shift"* (S7, WS2). Illustrating this point, other participants compared this moral shift with other possible outcomes from sabotage. One discussed an outcome in which sabotage successfully leads to a rolling back of the deployment targeted: *"It's obviously a good thing, but if it stops people from— if it makes people feel that the problem has gone away, I guess, maybe the problem we see is the way that tech companies interact with the city, or something like that. Then maybe it kills the resistance movement and makes it harder for us to resist other fights"* (S8, WS3). Here, S8 worries that without a collective moral shift, resistance can be pacified without deeper change. And another

saboteur emphasized the importance of this moral shift in order to avoid starting from scratch to resist every single unjust algorithmic system deployment: *"To me, I feel like the collective memory of the public's displeasure is best case because there's no end to these smart city deployments, so I think that kind of memory is a better safeguard than many of the other outcomes"* (S6, WS2). Additionally, S9 emphasized how this moral shift requires being involved in deeper, on-the-ground organizing and relationship building, rather than relying solely on sabotage: *"I think there's a need for these actions to be taken with some kind of neighborhood-level support or understanding or relationship building rather than, 'We're just coming in and destroying things'"* (S6, WS2).

In contrast, deployers largely considered responses to destruction that don't change the system's existing logics

In contrast to the aims of sabotage, responses to sabotage described by deployers failed to deeply engage with fundamentally problematic aspects of deployed algorithmic systems, instead preferring responses that mostly aimed to increase trust around largely unchanged existing deployments. In order to increase trust, deployers described choosing responses that were not explicitly antagonistic to saboteurs. For example, one deployer said that *"demonizing the attackers is just never a good strategy. ... you're trying to build trust in your company and in the systems that you build"* (D1, WS4). At the same time, deployers implicitly dismissed saboteurs' concerns about the nature of algorithmic system deployments by characterizing them as trust issues.

Deployers shared a variety of responses to algorithmic system sabotage that involve changing the information conveyed about the targeted algorithmic system. One method for this was to increase transparency around the algorithmic system and its interactions with people. For example, D6 talked about *"trying to have that transparency honestly somehow incorporated in this so that people can build trust. ... I work in the autonomous vehicle space, and being very transparent about the data, the safety data, the crash data, that's really important"* (WS6). Similarly, D2 said that sharing uses of camera-collected data with people whose data is collected helps them *"feel more comfortable being in front of the camera"* (WS4). Notably, increases in transparency described by deployers did not involve changing anything about the algorithmic system or its uses, just the information shared about the system and its uses.

Deployers also discussed how they could respond to algorithmic system sabotage through efforts to publicly emphasize the benefits of the targeted systems. One deployer talked about illustrating the positive impacts of delivery robots: *"Our robots helped grandma get her medicine without her having to exert herself and put herself in danger of falling. That's a good PR move; that could be a commercial"* (D5, WS6). Another similarly spoke of redirecting public framings of algorithmic cameras that monitor workers: *"Worker safety was the underlying principle that they were trying to go with, versus surveillance of misbehavior. I think they should have leaned into the worker safety principle a bit more and then communicated that, and built their messaging around it"* (D1, WS4). In PR responses like these, deployers aimed to valorize their deployed systems and bring more positive sentiment rather than backlash. D2 summarized, saying, *"Making sure that those good points are really what people are learning about all of these systems is the priority over some of the negative"* (WS4). Promoting the advantages of algorithmic systems, similar to increasing transparency, did not involve any modifications to the system deployments themselves, focusing instead on reshaping public sentiment in more favorable ways.

Another way deployers talked about changing PR strategies around targeted algorithmic systems entailed acknowledging fault for problematic system deployments. One participant emphasized the importance of accepting blame for building trust: *"Being honest with everyone about either where you went wrong as the company or where the product went wrong in your implementation of it is very important to build that trust back up"* (D2, WS4). Deployers also described improved deployments as ways to acknowledge issues. For instance, D1 said, *"There has to be a mea culpa of sorts, like, 'Okay, we messed up in the rollout of this system. We have taken XYZ steps to improve it, and we will roll it out in a more responsible, measured way'"* (WS4). And D3 described wanting to pause use of the system in order to *"acknowledge the moment and the concerns, and it allows the company to be thoughtful about how it would proceed. ... I think that could be interpreted as, 'We're just gonna pause it until the tumult dies down, and then we can just put it back on the market, and it'll be fine' so I could imagine the economists being persuaded to do it as well, but I would want to do it for what I would interpret as the ethical reasons"* (WS5). Here, D3 discusses navigating different incentives held by decision-makers that differ from their own preferences. In these examples, accepting blame can come with additional consideration of how system issues might have occurred and what to do about them.

To that end, deployers described responding to algorithmic system sabotage by further investigating system performance, as in section 9.4.1, and they aimed to use the information gained to deploy improved algorithmic systems that inspired greater trust. For instance, one participant said, *"At the end of the day, you want to build trust. ... And so I think the thing I would do, I'd like to shut the whole system down ... and troubleshoot what's wrong with that, and then try to deploy a better system"* (D1, WS4). D1 added that in addition to further investigating issues, testing an algorithmic system can show that it deserves trust, describing a process to *"rigorously test ... because that's how you build trust in these systems to show that they are actually working effectively"* (WS4).

However, deployers also talked about how they pursued further investigations in order to create the appearance of making changes without actually making changes. One deployer said, *"You'd like to do the minimum necessary to get everyone off your back, and to say that you've listened, and that you are still paying attention to them, but also they're gonna keep using the product, and they're gonna keep paying you for it"* (D4, WS5). As such, this deployer describes feedback as a way to make people feel heard without actually listening and changing anything meaningful about the deployment. Further emphasizing the lack of real interest in expansively collecting feedback, another deployer discussed potential implementation of a new way to collect and analyze feedback: *"I could imagine a company going this route, but only after they've already deployed a formal feedback mechanism. And they're gonna say, 'I don't care what people have emailed us up to this point, let's start fresh with our form that we control, where the messages are coming directly to us through this one pipeline'"* (D3, WS5). Here, the company feigns interest in further investigation of issues while actually limiting the feedback they engage with. D3 additionally noted that collecting feedback *"might deter attacks, because you're able to file a complaint"* (WS5), suggesting a belief that the aim of collecting feedback is placation of backlash rather than addressing problematic deployments.

Deployers did also mention some consideration of higher-level problem formulation in response to algorithmic system sabotage. Deployers described doing this by engaging with different levels of people involved in deployment processes. D2 stressed the necessity of going to the top level and *"having an honest conversation with*

whoever's in charge of deploying these systems, or purchasing them, or however it's structured, of whether it really is meeting that end goal" (WS4). At the same time, D4 stressed the opposite: *"You need to have it in a more bottom-up set up, and less of a top-down, forced way of doing things"* (WS5). Though they call for prioritizing engagement with different groups, together these examples highlight the importance of power dynamics in problem formulation processes.

Reconsidering problem formulation often went hand-in-hand with investigating issues further; that is, deployers occasionally discussed how investigation can be at broader levels in which problem formulation is reconsidered. For instance, one deployer said, *"You basically have to relaunch, and find out what the real problems are that your tech is missing. I don't like the idea of saying, 'We have a hammer, can you show us your nails?'"* (D3, WS5). To that end, another participant also emphasized the importance of reexamining what the real problems are: *"What is the user problem that's trying to be solved? What is the core of it? And attack that problem versus put a band-aid on whatever the current issues are"* (D1, WS4). In these examples deployers moved to a level where they described interest in actually reconceptualizing their algorithmic systems' uses and impacts.

9.5 Limitations

Our speculative scenarios were developed based on *observable* real-world cases; we assume that there are also many cases that go without notice, that are done with the intention of plausible deniability, that are designed to never be discovered, that happen internally in workplaces where actions are less likely to be made public. As such, we encourage future work to investigate these less visible forms of resistance and how saboteur and deployer views change when the observability changes. To that end, we conducted a speculative design study, so while we did learn a lot about participant motivations and perceptions, and did get to hear scaffolded discussions with conversations amongst multiple participants, we are unable to draw conclusions about what people would actually do in practice in situations like these. We encourage future work to examine how people decide their actions in real-world settings. Our participants were all English speakers and predominantly located in the United States (though one was in Argentina), and our recruitment was limited to those who would talk to us. This makes it more challenging to extend our findings to saboteurs who may have preferred to remain fully anonymous and to deployers who are not interested in revisiting moments of backlash to their deployed systems. There are opportunities for future work to investigate the views of these groups in particular, as well as to focus more on particular contexts of algorithmic sabotage.

9.6 Discussion

In this chapter, we investigated how AI sabotage is understood by both saboteurs and deployers of algorithmic systems, with the aim of informing ways for the HCI community to better support more productive engagement with algorithmic sabotage and resistance more broadly. The findings showcased significant divergences and alignments in saboteur and deployer perspectives, especially in how they conceived of what algorithmic sabotage is and which theories of change motivate it. Throughout our results, participants described various interpretations of algorithmic sabotage—as a communication practice, as an imaginative practice, and as an

ideological practice. Here, we discuss directions forward for how the HCI community can support more productive engagement with the phenomenon of algorithmic sabotage across these multiple levels, highlighting the complex nature of resistance to algorithmic systems as well as the challenges and opportunities that come with understanding and supporting it.

9.6.1 Supporting AI sabotage as communicative practice

The results of our study suggested that both saboteurs and deployers view acts of algorithmic sabotage as forms of communication. Deployer interpreted sabotage largely as “feedback” that can help them *improve algorithmic systems*, but saboteurs disagreed. Saboteurs rejected the feedback framing not because they are opposed to participating in feedback processes at all but because the AI feedback processes they have experienced have been broken to the extent that they seek sabotage as an alternate mode of recourse—and they are not the first to express this feeling [588]. Our results indicate that this stems in part from overly constrained communication channels between deployers and those resisting algorithmic systems that filter any feedback they receive into narrow conceptions of what change is possible. To that end, saboteurs might be more interested in feedback processes that are more imaginative about potential problems and solutions beyond aiming to reform existing technologies, and that provide saboteurs and resisters more control over the outcomes of the process.

For their part, saboteurs understood sabotage as communication toward different audiences and toward different ends than deployers did. In particular, saboteurs expressed wanting their acts to convey that resistance is happening to others who might then be spurred to resist. This framing parallels the concept of “bright spots”, or “observable community successes beyond the norm” in environmental conservation research [60]. That is, saboteurs see algorithmic sabotage as creating “bright spots” of visible success in resisting AI—and there are opportunities here for helping these spots shine even brighter by bringing more attention to them in both research and more public facing media. While ideally this attention would also legitimize these bright spots of resistance, our saboteur participants noted that even antagonistic or dismissive engagements with algorithmic sabotage can help make sites of contestation more visible and thus still meet some of their communication aims. That is, saboteurs described how algorithmic sabotage allows them to change the terrain that everyone operates on, often forcing deployers to communicate more widely and on saboteurs’ terms with their responses to algorithmic sabotage.

In that vein, saboteurs also wanted their acts to communicate dissatisfaction with algorithmic system deployments to society broadly, in order to develop a collective and long-term memory that can facilitate and sustain movements to resist various forms of injustice. How might acts of sabotage be supported to spread and coordinate movements? There is opportunity for the HCI community to continue examining historical and ongoing movements of tech resistance (e.g., [441]), such that people resisting algorithmic systems can more clearly understand their specific sociotechnical context and build solidarity and that deployers of algorithmic systems might themselves learn about and join movements of algorithmic resistance.

Similarly, both deployer and saboteur participants discussed desires for those affected by algorithmic systems to have more knowledge about those systems. However, the tenor of this knowledge varied; while people in both groups used terms like “AI literacy”, they had widely ranging interpretations of what this meant in practice

and why it was important. While deployers focused mainly on greater system transparency so that people could trust those systems more—spreading knowledge rather than changing anything about the system—saboteurs focused more on developing critical consciousness that can drive strategies for resistance and repair: spreading knowledge in order to change the system. This contrast indicates the need for further reflection on the extent to which directions like algorithmic literacy are helpful for equipping people with the information needed for resistance, versus contribute to ossification of mindsets that algorithmic systems are inevitable and that individuals' experiences with AI alone do not bring them enough expertise to take action. How do methods in areas like algorithmic literacy compare to on-the-ground practices of mutual aid, support, and skill sharing? While we see opportunities for HCI researchers to further support on-the-ground practices of information sharing, following [524] we do caution that some details and tactics should not be shared in academic venues as this can lead to attempts to frustrate rather than support algorithmic resisters.

9.6.2 Supporting AI sabotage as design practice

Our findings also indicate the imaginative potential of algorithmic sabotage. This is in line with Ruha Benjamin's conception of imagination as a way to break down and move beyond structures of tyranny [57]. To that end, saboteurs described how they viewed algorithmic sabotage as a generative practice, a practice of design itself. This follows work arguing that acts of refusal to technologies, like algorithmic sabotage, can be understood as acts of design [596], as well as work that considers how acts of dismantling sociotechnical systems, of taking them apart, are also acts of *remantling*: of collaboratively creating and constructing new realities [52].

Here we see opportunities for the HCI community to work to further understand in what ways algorithmic sabotage on the ground embodies and contributes to design principles. For example, in our study the materiality of an algorithmic system seemed to matter, as physical instantiations provide different sorts of affordances and resistance possibilities than fully digital instantiations of algorithmic systems. Participants discussed different sorts of risks and expertise needed depending on these differences in system materiality—such as a smart lamppost allowing spray paint or throwing rocks but requiring risks of acting in public, whereas drone software requiring code-writing prowess but might allow plausible deniability if caught. These different skills needed, tactics selected, and mindsets around risks accepted lead to distinctly different forms of designing new sociotechnical systems through sabotaging AI—opening questions around how to expansively consider the characteristics that constitute a designer while upholding the value of expertise.

Algorithmic sabotage can also imaginatively move toward social justice in ways that extend beyond existing categories of design and making as a form of “elimination for good”, an emancipatory act of unmaking [441]. In this vein, there are implications for HCI to support deployers of algorithmic systems to reconceive of their plans for building and launching their systems such that they include ways to not only adjust but discontinue the deployment, to *undesign* technologies [399]. In the realm of algorithmic systems, this includes forms of retraction such as algorithm abandonment [280], as well as more forceful methods of direct action that involve obstruction or destruction of algorithmic systems.

Algorithmic sabotage offers unique lessons for the HCI community. We see opportunities to support deployers to themselves participate in algorithmic sabotage. Though we focused on two distinct groups of saboteurs and deployers, these groups

can have overlap: a number of our saboteur participants had at times also been deployers of algorithmic systems. This might involve finding ways to bring more visibility to acts of sabotage and other forms of resistance to algorithmic systems that are already happening from deployers themselves, as these actions tend to go more unnoticed. Any methods that aim to do this must also balance the reality that deployers committing algorithmic sabotage may rely on plausible deniability or an ability to not be found out, which would make them naturally wary of connecting with others and want to evade publicity.

We also see opportunities for the design of affordances in algorithmic systems that are specifically intended for breaking or blocking those systems. This parallels related notions of un-crafting [375], unmaking [491, 492], obsolescence and breakdown [273] as relevant to design practices and ways to provide space for more expansive modes of engagement with sociotechnical systems. Built-in affordances for sabotage can be understood as a way to ensure ongoing consent from those who interact with algorithmic systems. And designing for algorithmic sabotage provides a way for deployers to intentionally probe for intensely negative reactions to their algorithmic systems that may otherwise have been unheard, underappreciated, or underexpressed while also allowing those reactions to avoid broken feedback channels and have direct impact immediately.

9.6.3 Supporting AI sabotage as ideological practice

Throughout our study's results, participants discussed algorithmic sabotage at an ideological level, especially discussing how algorithmic sabotage attends to issues of and aims related to power. Though there were many disconnects between saboteur and deployer perceptions of algorithmic sabotage, saboteur views of resistance did resonate with deployers at certain points: both viewed sabotage as a way to attend to lacking power and control, though deployers largely considered this within the contexts of privacy and work. Additionally, some deployers described desires to reconsider the higher-level problem formulation of algorithmic systems, in line with saboteur desires to resist not just specific instantiations of algorithmic systems but the larger political projects of algorithmic systems.

These points of resonance between saboteurs and deployers of algorithmic systems suggest opportunities to support bridge-building between these groups. We see openings to help deployers broaden their conceptions of where power and control is relevant, such as by helping develop their critical consciousness around the ideological and political project of AI through which power is consolidated and shifted away from individuals and toward owners of algorithmic systems [20, 506, 545].

Many deployers are themselves positioned closer to the individuals with little power than to the AI owners, which might prime them to be more sympathetic to those who engage in algorithmic sabotage and other forms of resistance even as they feel constrained by their own workplace incentive structures and internal hierarchies. In particular, the deployers in our study did not necessarily hold higher positions of decision-making power and expressed some stances related to ethics and justice that paralleled those of our saboteur participants. For example, D3 described how an AI system's problem formulation was entirely incorrect and should

be reconceptualized; at the same time, D3 also described needing to frame any actions in ways that resonate with other decision-makers and their incentives. Following research that has identified opportunities for tech workers in similar positions, we see opportunities for HCI to support “soft” tactics of resistance that operate within a larger, dominant technology industry. This might involve the creation of tools designed specifically for deployers in positions like this to help them act in ways sympathetic to saboteurs by framing their actions in ways palatable to values within their organizations.

Additionally, we see opportunities to support more overt resistance from deployers, such as designing new strategies for building solidarity through formal and informal workplace unions. This space requires new methods to help deployers themselves avoid, resist, and sabotage algorithmic systems that their organizations turn on them to suppress efforts for collective power [424]. To that end, there are opportunities for recruiting deployers as saboteurs, building bridges between those who design, develop, and launch algorithmic systems and those who are affected by and engage in on-the-ground resistance of algorithmic systems. It is already the case that some algorithmic system deployers advocate for people impacted negatively by their systems (e.g., [141]). At the same time, given existing feedback processes, saboteurs are not interested in algorithmic sabotage being used for design feedback. As such, we see opportunities to connect advocating deployers with sabotage and resistance movements—in safe ways that do not reveal identities of either—so that deployers might further develop their critical consciousness, share information about points of leverage regarding the algorithmic systems they deploy, and begin participating in algorithmic sabotage themselves from the inside.

9.7 Conclusion

We provided a deeper, more nuanced understanding of saboteurs’ and deployers’ views of algorithmic sabotage and its aims, highlighting significant points of disconnect and partial alignment. Our findings indicate that 1) saboteurs and deployers both view algorithmic sabotage as a form of communication, but for different audiences and purposes and 2) saboteurs view algorithmic sabotage as an imaginative practice, while deployers write off algorithmic sabotage as unimaginative. These findings indicate opportunities for the HCI community to build bridges between these saboteurs and deployers in order to support broader movements of resistance. For example, needs for AI literacy can be interpreted as a need for further development of critical consciousness, encouraging deployers to share knowledge of and skills related to AI systems specifically for the purposes of more effective resistance. Additionally, algorithmic systems might be intentionally built with affordances for sabotage that can be used by both saboteurs or deployers if the system acts problematically. We highlight that power dynamics mediate not only acts of sabotage but also deployer responses to those acts, so supports for either group will require a careful consideration of specific hierarchies and risks involved.

Chapter 10

Conclusion

This dissertation has investigated a range of acts of bottom-up resistance to harmful AI systems—from everyday algorithm auditing to algorithmic sabotage, from individual workarounds to collective pressure, from communicating issues upward to communicating contestation outward—highlighting the breadth of tactics that people use to oppose harmful AI systems. Throughout, bottom-up resistance has revealed itself as an embodied practice that creates material change in the world. Even when structurally limited, people have some power and take action, leveraging it to begin designing the worlds they want. This dissertation emphasizes the importance of affected people and their self-determined actions in AI harm remediation efforts and in sociotechnical spaces more broadly, highlighting the capabilities of bottom-up resistance to provide an antidote to insufficient existing mechanisms for accountability and redress around harmful AI systems.

I understand this research as an attempt to bring more focus in research and technological spaces to acts of bottom-up resistance that have been happening, are happening, and will continue to happen. My contributions then fall largely into two categories: understanding the details and depths of how bottom-up resistance is enacted on the ground, then translating that into research spaces while synthesizing arguments for why these acts deserve space and legitimacy as well as implications for what researchers and technologists, once convinced, can do next. Throughout I have explored the dynamics of power, the complexities of pushing for change, and the nuances of positionality—but much remains to be explored on these fronts. To that end, building on this dissertation, here I rhapsodize on directions for future work:

Moving forward, this work can be extended by **developing theoretical frameworks** that capture the situated, evolving, and heterogeneous ways that resistance unfolds. Different groups of people have vastly different needs and motivations for their engagements with and resistance to harmful AI systems. Additionally, as the implementations and features of AI systems change over time, different forms of resistance become more or less tenable. As such, future work should continue to study AI resistance in a variety of contexts in order to provide detailed, nuanced, empirical knowledge of what is happening on the ground around AI systems and what opportunities exist to support affected people's self-determination moving forward.

Working to understand and support acts of bottom-up resistance also begs **more reflection regarding risks** from multiple angles. For one, the risks of resisting differ depending on the context. Throughout my work, the importance of knowledge sharing came up many times as a way to form collectives, to share solidarity, to understand and resist systems more effectively, to develop critical consciousness. How can knowledge about the complexities of risks in these processes be developed, recorded, and shared to support resisters in situation-specific risk analysis, so that they can develop better precautions to limit their risks and decide exactly what

sorts of safety they are willing to cede when acting? Sharing information about bottom-up resistance can itself come with risks. How can we ensure that research and tech spaces direct more attention to supporting resistance efforts *without overstepping* by making sensitive details public? And supporting bottom-up resistance can also come with risks of supporting bad actors, as not all bottom-up actions are positive. How can supports be designed such that they are more difficult to redirect in harmful ways themselves? How can we both maintain the importance of expertise for action while expanding notions around what expertise is?

To that end, there is significant contention around what is a “good” versus a “bad” act of resistance. Future research should work to **further understand the nuances around forms of legitimacy** that bottom-up resistance to AI both attempts to work within and break free from. For example, everyday algorithm auditing involves people collecting information and conveying issues within frames of logic, quantification, and proof held by decision-makers; at the same time, everyday algorithm auditing can also involve pushing collectively for change in more forceful ways that do not bow to acceptability standards of the powerful. Here, emotion and use offer two particularly interesting points of entry. Negative emotions like anger are often used as reasons to dismiss forms of bottom-up resistance like AI sabotage as unreasonable; however, anger can provide significant push to drive individual resisters and larger movements of resistance forward. Additionally, acts of bottom-up resistance to AI are often written off as misuses of AI systems, even though—as shown by my work—expansions on the category of use when people resist and attempt to reconfigure AI systems are important for meaningfully addressing AI harm. How can we move away from limiting ideologies of legitimacy and toward a vision that centers community agency, resistance, and transformative change to inform the governance of AI?

Appendix A

Ch. 4 Demographic information

	#	%		#	%		#	%
Gender			Race/Ethnicity			Political Views		
Female	14	61%	White	8	35%	Very Liberal	7	30%
Male	7	30%	Asian	7	30%	Liberal	6	26%
Non-binary	2	9%	Black or African American	6	26%	Moderate	5	22%
Sexual Orientation			American Indian or Alaska Native	2	9%	Conservative	3	13%
Straight	15	65%	Hispanic or Latino	2	9%	Prefer not to disclose	2	9%
Bisexual	4	17%	Middle Eastern	1	4%	Highest Degree Completed		
Lesbian	2	9%	Religion			Bachelor's degree	7	30%
Gay	1	4%	Protestant	5	22%	Some college	5	22%
Queer	1	4%	Nothing in particular	4	17%	Master's/professional degree	5	22%
			Hindu	3	13%	Doctorate	2	9%
			Atheist	3	13%	Some grad school	2	9%
			Something else	3	13%	Highschool/equivalent	2	9%
			Roman Catholic	2	9%	Income		
			Prefer not to disclose	2	9%	\$75–100k	6	26%
			Buddhist	1	4%	\$50–75k	6	26%
			Muslim	1	4%	\$30–50k	6	26%
						\$15–30k	3	13%
						<\$15k	1	4%
						Prefer not to disclose	1	4%

FIGURE A.1: Self-reported participant demographic information from the screening questionnaire.

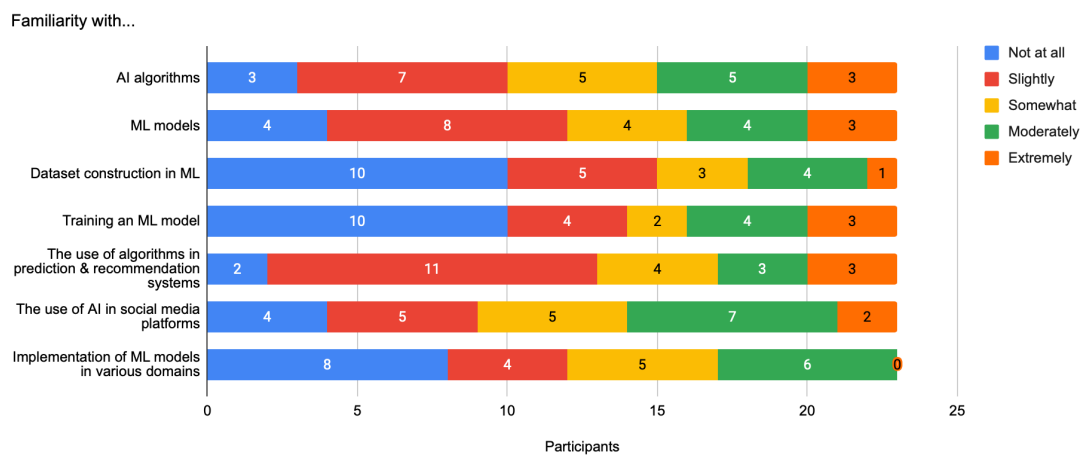


FIGURE A.2: Participant familiarities with various algorithmic concepts, on a 5-point Likert scale.

Appendix B

Ch. 7 Additional Details on Process of Response Collection & Categorization

To explore the space, we adopted a case study approach, gathering documented historical cases of responses from below that intend to remediate perceived algorithmic harm. We understand actors and actions from below as those lacking direct power to make decisions about the design and deployment of algorithmic systems. To emphasize power relations, we broadly refer to those who have control over algorithmic systems as “*algorithmlords*”, building on past work describing “data barons” who control data and “digital feudalism” in which tech companies behave as modern-day feudal lords [159, 205, 275]. In the vein of [568], we understand *algorithmlords* as those holding greater power to envision and construct algorithmic systems, occupying positions “above” those subjugated by their decisions, while acknowledging that relationships and interactions between people are more complex and nuanced than this fully captures. We included cases in which the system is described as algorithmic in the source material (making no attempt to further define what an algorithm is [339]), the actor perceives harm from that system and takes action in response, and the actor lacks direct power to make decisions about the design or deployment of algorithmic systems (following [568]).

We began with a set of 10 documented cases drawn from seven past research papers that the authors were familiar with [148, 168, 169, 288, 302, 474, 552]. From our initial set of research papers, we used both forward and backward citation chaining to find additional examples, following relevant citations to find other cases. We also conducted searches of research literature, news media, and social media using combinations of relevant keywords (e.g., “algorithmic”, “harm”, “users”, “respond”) to find additional relevant cases. Our cases spanned many different algorithmic domains (e.g., recommendation, image cropping, facial recognition) and types of harm (e.g., financial, stereotyping, surveillance).

While collecting cases, we also began analyzing and categorizing the responses present in the cases we had already gathered. We coded each case with the algorithmic harm identified by the responder(s), who was impacted, the algorithmic system, the actor(s) responding, the response action or actions, any tools used, and if described, the desired and actual outcomes. As a single case often included responding in multiple ways, we extracted all the responses from each case. We conducted a bottom-up thematic analysis with a series of interpretation sessions on these responses and, based on this, grouped our set of responses into *types* of responses based on similar responder behaviors present, which we then synthesized into three high-level *strategies*, each of which captures a different way that these responses aim

to interact with power.

Finally, based on discussions with researchers beyond our team, we identified additional relevant cases that could inform the development and refinement of our taxonomy. For example, Global North locations were overrepresented in our initial set of cases. To mitigate this bias, we added specific responses from cases that emerged in our discussions and from conducting additional searches that covered other regions.

We conducted initial searches in the spring of 2022, then coded and interpreted as already described in the Appendix. We then conducted additional searches that targeted regions underrepresented in our dataset by appending region names to the previous search terms and identified an additional 50 responses across 26 cases, using the criteria described above. We iteratively integrated the cases into our categories, expanding one of our categories in the process. Continuing our searches in 2023, we identified 19 more responses over 11 cases that, when coded and integrated in our analysis, did not impact the taxonomy's structure.

In all, we collected 96 cases from 75 sources with 169 extracted responses from below to algorithmic harm. We stopped collecting cases when we reached saturation in our categorization: that is, when we stopped finding examples of new types of responses. We emphasize that we see the cases used as the basis for our taxonomy as illustrative of the phenomenon under study, but not necessarily exhaustive. Additionally, due to the constantly changing and emerging nature of algorithmic systems, of their applications, of related harms, and of related responses to these harms, we see this work as an initial exploration and invite future researchers to expand on it.

Bibliography

- [1] uncle nole / dhara. @onyowalkman: “help does twitter only focus on the white dude??” with related images. 2020. URL: <https://twitter.com/onyowalkman/status/1308035295095316481>.
- [2] A-levels and gcse: How did the exam algorithm work? 2020. URL: <https://www.bbc.com/news/explainers-53807730>.
- [3] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. “Roles for computing in social change”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 252–260.
- [4] About. URL: <https://incidentdatabase.ai/about/>.
- [5] About the AIAAIC Repository. URL: <https://www.aiaaic.org/aiaaic-repository/about-the-aiaaic-repository>.
- [6] ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT). URL: <https://facctconference.org/>.
- [7] William Agnew. “What Can AI Ethics Learn from Anarchism?” In: *XRDS* 30.4 (June 2024), 22–25. ISSN: 1528-4972. DOI: 10.1145/3665594. URL: <https://doi.org/10.1145/3665594>.
- [8] Parag Agrawal and Dantley Davis. “Transparency around image cropping and changes to come”. In: *Twitter* (2020). URL: https://blog.twitter.com/en_us/topics/product/2020/transparency-image-cropping.html.
- [9] Alex Ahmed. “We Will Not Be Pacified Through Participation”. In: *Proceedings of the CSCW 2020 Workshop on Collective Organizing and Social Responsibility*. Tech Otherwise, 2020. DOI: 10.21428/93b2c832.e4aa9a47. URL: <https://techotherwise.pubpub.org/pub/qanxilt9/release/1>.
- [10] Sara Ahmed. *Complaint!* New York, USA: Duke University Press, 2021. ISBN: 9781478022336. DOI: doi:10.1515/9781478022336. URL: <https://doi.org/10.1515/9781478022336>.
- [11] Sara Ahmed. “Feminist Attachments”. In: *The Cultural Politics of Emotion*. NED - New edition, 2. Edinburgh University Press, 2014, pp. 168–190. ISBN: 9780748691135. URL: <http://www.jstor.org/stable/10.3366/j.ctt1g09x4q.13> (visited on 01/22/2025).
- [12] Sara Ahmed. No. 2017. URL: <https://feministkilljoys.com/2017/06/30/no/>.
- [13] IBM Research Trusted AI. “AIF360 API”. In: (2021). URL: <https://aif360.mybluemix.net/>.
- [14] IBM Research Trusted AI. “AIX360 API”. In: (2021). URL: <https://aix360.mybluemix.net/>.
- [15] Open AI. *ChatGPT Feedback Contest: Official Rules*. 2022. URL: https://cdn.openai.com/chatgpt/ChatGPT_Feedback_Contest_Rules.pdf.

- [16] Organizers of Queer in AI, Nathan Dennler, Anaelia Ovalle, Ashwin Singh, Luca Soldaini, Arjun Subramonian, Huy Tu, William Agnew, Avijit Ghosh, Kyra Yee, Irene Font Peradejordi, Zeerak Talat, Mayra Russo, and Jess De Jesus De Pinho Pinhal. "Bound by the Bounty: Collaboratively Shaping Evaluation Processes for Queer AI Harms". In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '23. Montréal, QC, Canada: Association for Computing Machinery, 2023, 375–386. ISBN: 9798400702310. DOI: 10.1145/3600211.3604682. URL: <https://doi.org/10.1145/3600211.3604682>.
- [17] W.M. Akers. *Comrades: A Revolutionary RPG*. 2019. URL: <http://wmakers.net/comrades>.
- [18] Leigh Alexander. *Do Google's 'unprofessional hair' results show it is racist?* 2016. URL: <https://www.theguardian.com/technology/2016/apr/08/does-google-unprofessional-hair-results-prove-algorithms-racist->.
- [19] AlgorithmWatch. *Can you break the algorithm?* 2023. URL: <https://algorithmwatch.org/en/break-algorithm-game>.
- [20] Ali Alkhatib. *Defining AI*. 2024. URL: <https://ali-alkhatib.com/blog/defining-ai>.
- [21] Ali Alkhatib. *Destroy AI*. 2024. URL: <https://ali-alkhatib.com/blog/fuck-up-ai>.
- [22] 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*. CHI '21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445740. URL: <https://doi.org/10.1145/3411764.3445740>.
- [23] Amy Allen. "Feminist Perspectives on Power". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Fall 2022. Metaphysics Research Lab, Stanford University, 2022.
- [24] Heather Allen. *Activists in San Francisco demand governor shut down self-driving cars*. 2024. URL: <https://www.nbcbayarea.com/news/local/san-francisco-protest-waymo-cruise-cars/3458525/>.
- [25] Micah Altman, Alexandra Wood, and Effy Vayena. "A Harm-Reduction Framework for Algorithmic Fairness". In: *IEEE Security & Privacy* 16.3 (2018), pp. 34–45. DOI: 10.1109/MSP.2018.2701149.
- [26] ameliorate_d. @ameliorate_d: "Trying a horrible experiment." with canine images. 2020. URL: https://twitter.com/ameliorate_d/status/1307576187942703104.
- [27] Dimitra Anastasiou and Rajat Gupta. "Comparison of crowdsourcing translation with Machine Translation". In: *Journal of Information Science* 37.6 (2011), pp. 637–659.
- [28] Nazanin Andalibi, Oliver L. Haimson, Munmun De Choudhury, and Andrea Forte. "Social Support, Reciprocity, and Anonymity in Responses to Sexual Abuse Disclosures on Social Media". In: *ACM Trans. Comput.-Hum. Interact.* 25.5 (2018). ISSN: 1073-0516. DOI: 10.1145/3234942. URL: <https://doi.org/10.1145/3234942>.
- [29] Michael Anderson and Jeremy Magruder. "Learning from the crowd: Regression discontinuity estimates of the effects of an online review database". In: *The Economic Journal* 122.563 (2012), pp. 957–989.

- [30] Markus Angermeier. @kosmar: two identical images with different bright colored backgrounds in a series of tweets using different colors. 2020. URL: <https://twitter.com/kosmar/status/1307777202445004800>.
- [31] Markus Angermeier. @kosmar: two images for comparison with different skin tones. 2020. URL: <https://twitter.com/kosmar/status/1307776710604185600>.
- [32] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. *Machine Bias*. 2016. URL: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- [33] *Another wrongful death lawsuit from C.A. Goldberg, PLLC against Amazon.com, Inc.* 2022. URL: <https://www.cagoldberglaw.com/wrongful-death-plaintiffs-vs-amazon-com-inc-c-a-goldberg/>.
- [34] Tony Arcieri. @bascul: "Trying a horrible experiment... Which will the Twitter algorithm pick: Mitch McConnell or Barack Obama?" with related images. 2020. URL: <https://twitter.com/bascul/status/1307440596668182528>.
- [35] Crosby Armstrong. *Video Games Remain America's Favorite Pastime With More Than 212 Million Americans Playing Regularly*. 2023. URL: <https://www.theesacom/video-games-remain-americas-favorite-pastime-with-more-than-212-million-americans-playing-regularly>.
- [36] Sherry R. Arnstein. "A Ladder Of Citizen Participation". In: *Journal of the American Institute of Planners* 35.4 (1969), pp. 216–224. DOI: 10.1080/01944366908977225. eprint: <https://doi.org/10.1080/01944366908977225>. URL: <https://doi.org/10.1080/01944366908977225>.
- [37] Imanol Arrieta-Ibarra, Leonard Goff, Diego Jiménez-Hernández, Jaron Lanier, and E. Glen Weyl. "Should We Treat Data as Labor? Moving beyond "Free"". In: *AEA Papers and Proceedings* 108 (2018), 38–42. DOI: 10.1257/pandp.20181003. URL: <https://www.aeaweb.org/articles?id=10.1257/pandp.20181003>.
- [38] Artefact. *The Most Likely Machine: Helping children develop algorithmic literacy*. URL: <https://www.artefactgroup.com/resources/most-likely-machine>.
- [39] Avi Asher-Schapiro. INTERVIEW - 'Algorithms of oppression': Big tech urged to combat discrimination. 2020. URL: <https://news.trust.org/item/20201111000521-t3k8r>.
- [40] Joshua Asplund, Motahhare Eslami, Hari Sundaram, Christian Sandvig, and Karrie Karahalios. "Auditing Race and Gender Discrimination in Online Housing Markets". In: *Proceedings of the International AAAI Conference on Web and Social Media* 14.1 (2020), pp. 24–35. URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/7276>.
- [41] Blair Attard-Frost. "Transfeminist AI governance". In: *First Monday* 30.4 (2025). DOI: 10.5210/fm.v30i4.14121. URL: <https://firstmonday.org/ojs/index.php/fm/article/view/14121>.
- [42] Josh Attenberg, Panagiotis G. Ipeirotis, and Foster Provost. "Beat the machine: challenging workers to find the unknown unknowns". In: *Proceedings of the 11th AAAI Conference on Human Computation*. AAAIWS'11-11. AAAI Press, 2011, 2–7.
- [43] Joshua Attenberg, Panos Ipeirotis, and Foster Provost. "Beat the Machine: Challenging Humans to Find a Predictive Model's "Unknown Unknowns"". In: *J. Data and Information Quality* 6.1 (Mar. 2015). ISSN: 1936-1955. DOI: 10.1145/2700832. URL: <https://doi.org/10.1145/2700832>.

- [44] Garrett B. *Very bothered by Yelp Review Filter Part 1*. 2011. URL: <https://www.yelp.com/topic/los-angeles-very-bothered-by-yelp-review-filter-part-1>.
- [45] Avraham Yosef Baez. *Trapped – a roleplaying game remembering the Triangle Shirtwaist Factory*. URL: <https://avryos.itch.io/trapped>.
- [46] Alison Bailey. “On Anger, Silence, and Epistemic Injustice”. In: *Royal Institute of Philosophy Supplement* 84 (2018), 93–115. DOI: [10.1017/S1358246118000565](https://doi.org/10.1017/S1358246118000565).
- [47] Jack Bandy. “Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits”. In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW1 (Apr. 2021). DOI: [10.1145/3449148](https://doi.org/10.1145/3449148). URL: <https://doi.org/10.1145/3449148>.
- [48] Gagan Bansal and Daniel Weld. “A coverage-based utility model for identifying unknown unknowns”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.
- [49] Chelsea Barabas. “Refusal in Data Ethics: Re-Imagining the Code Beneath the Code of Computation in the Carceral State”. In: *Engaging Science, Technology, and Society* 8.2 (Sept. 2022), p. 23. DOI: [10.17351/ests2022.1233](https://doi.org/10.17351/ests2022.1233).
- [50] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. “The problem with bias: Allocative versus representational harms in machine learning”. In: *9th Annual conference of the special interest group for computing, information and society*. 2017.
- [51] Solon Barocas and Andrew D Selbst. “Big data’s disparate impact”. In: *Calif. L. Rev.* 104 (2016), p. 671.
- [52] Eric P. S. Baumer and Vera Khovanskaya. “Sociotechnical Remantling”. In: *Proceedings of the Sixth Decennial Aarhus Conference: Computing X Crisis*. AAR ’25. New York, NY, USA: Association for Computing Machinery, 2025, 27–41. ISBN: 9798400720031. DOI: [10.1145/3744169.3744170](https://doi.org/10.1145/3744169.3744170). URL: <https://doi.org/10.1145/3744169.3744170>.
- [53] Eric P.S. Baumer and M. Six Silberman. “When the Implication is Not to Design (Technology)”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’11. Vancouver, BC, Canada: Association for Computing Machinery, 2011, 2271–2274. ISBN: 9781450302289. DOI: [10.1145/1978942.1979275](https://doi.org/10.1145/1978942.1979275). URL: <https://doi.org/10.1145/1978942.1979275>.
- [54] Venture Beat. @VentureBeat: “AI predictions for 2019 from Yann LeCun, Hilary Mason, Andrew Ng, and Rumman Chowdhury <https://wp.me/p8wLEc-ahoT> by @khar-ijohnson”. 2019. URL: <https://twitter.com/VentureBeat/status/1080485485473083392>.
- [55] Ashley Belanger. *AI haters build tarpits to trap and trick AI scrapers that ignore robots.txt*. 2025. URL: <https://arstechnica.com/tech-policy/2025/01/ai-haters-build-tarpits-to-trap-and-trick-ai-scrapers-that-ignore-robots-txt/>.
- [56] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. “AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias”. In: *IBM Journal of Research and Development* 63.4/5 (2019), pp. 4–1.
- [57] Ruha Benjamin. *Imagination: A Manifesto*. New York, USA: W. W. Norton & Company, 2024. ISBN: 9781324020974.

- [58] Ruha Benjamin. "Informed Refusal: Toward a Justice-based Bioethics". In: *Science, Technology, & Human Values* 41.6 (2016), pp. 967–990. DOI: [10.1177/0162243916656059](https://doi.org/10.1177/0162243916656059). eprint: <https://doi.org/10.1177/0162243916656059>. URL: <https://doi.org/10.1177/0162243916656059>.
- [59] Cynthia Bennett, Emily Ackerman, Bonnie Fan, Jeffrey Bigham, Patrick Carington, and Sarah Fox. "Accessibility and The Crowded Sidewalk: Micro-mobility's Impact on Public Space". In: *Proceedings of the 2021 ACM Designing Interactive Systems Conference*. DIS '21. Virtual Event, USA: Association for Computing Machinery, 2021, 365–380. ISBN: 9781450384766. DOI: [10.1145/3461778.3462065](https://doi.org/10.1145/3461778.3462065). URL: <https://doi.org/10.1145/3461778.3462065>.
- [60] Elena Bennett, Martin Solan, Reinette Biggs, Timon McPhearson, Albert Norström, Per Olsson, Laura Pereira, Garry Peterson, Ciara Raudsepp-Hearne, Frank Biermann, Stephen Carpenter, Erle Ellis, Tanja Hichert, Victor Galaz, Myanna Lahsen, Manjana Milkoreit, Berta Martin-López, Kimberly Nicolas, Rika Preiser, Gaia Vince, Joost Vervoort, and Jianchu Xu. "Bright Spots: Seeds of a good Anthropocene". English. In: *Frontiers in Ecology and the Environment* 14.8 (2016), pp. 441–448. ISSN: 1540-9295. DOI: [10.1002/fee.1309](https://doi.org/10.1002/fee.1309).
- [61] W. Lance Bennett and Alexandra Segerberg. "THE LOGIC OF CONNECTIVE ACTION". In: *Information, Communication & Society* 15.5 (2012), pp. 739–768. DOI: [10.1080/1369118X.2012.670661](https://doi.org/10.1080/1369118X.2012.670661). eprint: <https://doi.org/10.1080/1369118X.2012.670661>. URL: <https://doi.org/10.1080/1369118X.2012.670661>.
- [62] Niels van Berkel, Jorge Goncalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M. Kelly, and Vassilis Kostakos. "Crowdsourcing Perceptions of Fair Predictors for Machine Learning: A Recidivism Case Study". In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). DOI: [10.1145/3359130](https://doi.org/10.1145/3359130). URL: <https://doi.org/10.1145/3359130>.
- [63] Helen L. Berry, Thomas D. Waite, Keith B. G. Dear, Anthony G. Capon, and Virginia Murray. "The case for systems thinking about climate change and mental health". In: *Nature Climate Change* 8.4 (2018), pp. 282–290. DOI: [10.1038/s41558-018-0102-4](https://doi.org/10.1038/s41558-018-0102-4). URL: https://ideas.repec.org/a/nat/natcli/v8y2018i4d10.1038_s41558-018-0102-4.html.
- [64] Elettra Bietti. "From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Barcelona, Spain: Association for Computing Machinery, 2020, 210–219. ISBN: 9781450369367. DOI: [10.1145/3351095.3372860](https://doi.org/10.1145/3351095.3372860). URL: <https://doi.org/10.1145/3351095.3372860>.
- [65] Jeffrey P Bigham, Michael S Bernstein, and Eytan Adar. "Human-computer interaction and collective intelligence". In: *Handbook of collective intelligence* 57 (2015).
- [66] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. "'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions". In: *CoRR abs/1801.10408* (2018). arXiv: [1801.10408](https://arxiv.org/abs/1801.10408). URL: [http://arxiv.org/abs/1801.10408](https://arxiv.org/abs/1801.10408).
- [67] Sarah Bird. *Fairlearn API*. 2020. URL: https://fairlearn.github.io/v0.5.0/api_reference/fairlearn.datasets.html.

- [68] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. *Fairlearn: A toolkit for assessing and improving fairness in AI*. Tech. rep. MSR-TR-2020-32. Microsoft, 2020. URL: <https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/>.
- [69] Abeba Birhane. “Algorithmic injustice: a relational ethics approach”. In: *Patterns* 2.2 (2021), p. 100205. ISSN: 2666-3899. DOI: <https://doi.org/10.1016/j.patter.2021.100205>. URL: <https://www.sciencedirect.com/science/article/pii/S2666389921000155>.
- [70] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. “Power to the People? Opportunities and Challenges for Participatory AI”. In: *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. EAAMO ’22. Arlington, VA, USA: Association for Computing Machinery, 2022. ISBN: 9781450394772. DOI: [10.1145/3551624.3555290](https://doi.org/10.1145/3551624.3555290). URL: <https://doi.org/10.1145/3551624.3555290>.
- [71] Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. “The Forgotten Margins of AI Ethics”. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, 948–958. ISBN: 9781450393522. DOI: [10.1145/3531146.3533157](https://doi.org/10.1145/3531146.3533157). URL: <https://doi.org/10.1145/3531146.3533157>.
- [72] Abeba Birhane, Ryan Steed, Victor Ojewale, Briana Vecchione, and Inioluwa Deborah Raji. *AI auditing: The Broken Bus on the Road to AI Accountability*. 2024. arXiv: [2401.14462](https://arxiv.org/abs/2401.14462) [cs.CY]. URL: <https://arxiv.org/abs/2401.14462>.
- [73] Todd Bishop. “Amazon changes its key formula for calculating product ratings and displaying reviews”. In: *GeekWire* (June 2015). URL: <https://www.geekwire.com/2015/amazon-changes-its-influential-formula-for-calculating-product-ratings>.
- [74] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. “Language (Technology) is Power: A Critical Survey of “Bias” in NLP”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, July 2020, pp. 5454–5476. DOI: [10.18653/v1/2020.acl-main.485](https://doi.org/10.18653/v1/2020.acl-main.485). URL: <https://aclanthology.org/2020.acl-main.485>.
- [75] Esther Bond. *Google Fixes Gender Bias in Google Translate (Again)*. 2020. URL: <https://slator.com/google-fixes-gender-bias-in-google-translate-again/>.
- [76] Tiziano Bonini and Emiliano Treré. *Algorithms of Resistance: The Everyday Fight against Platform Power*. The MIT Press, Feb. 2024. ISBN: 9780262377485. DOI: [10.7551/mitpress/14329.001.0001](https://doi.org/10.7551/mitpress/14329.001.0001). URL: <https://doi.org/10.7551/mitpress/14329.001.0001>.
- [77] Masha Borak. *China’s Gig Workers are Challenging Their Algorithmic Bosses*. 2022. URL: <https://www.wired.com/story/chinas-gig-workers-challenging-algorithmic-bosses/>.

- [78] Joyann Boyce. @joyannboyce: "The following tweets are part of a Twitter algorithm experiment using Fenty foundation models." with related images. 2020. URL: <https://twitter.com/joyannboyce/status/1308080366369021954>.
- [79] Brenda Brathwaite and Ian Schreiber. *Challenges for Game Designers*. 1st ed. USA: Charles River Media, Inc., 2008. ISBN: 158450580X.
- [80] Virginia Braun and Victoria Clarke. "Reflecting on reflexive thematic analysis". In: *Qualitative research in sport, exercise and health* 11.4 (2019), pp. 589–597.
- [81] Rucksack Brian. *Major changes to review scoring method on Booking com*. 2019. URL: <https://hostelmanagement.com/forums/major-changes-review-scoring-method-booking-com.html>.
- [82] Ben Brody. *Weight Watchers must delete algorithms built from kids' data*. 2022. URL: <https://www.protocol.com/bulletins/weight-watchers-coppa-ftc>.
- [83] Rick Broida. *EA turns lemons into Lemonade Tycoon for iPhone*. 2009. URL: <https://www.cnet.com/tech/mobile/ea-turns-lemons-into-lemonade-tycoon-for-iphone/>.
- [84] Christopher Brooks, Erica Principe Principe Cruz, Jamie Camera, and Alexandra To. "Uplifting Us: A BIPOC Game Design Case Study". In: *Proceedings of the 16th International Conference on the Foundations of Digital Games*. FDG '21. Montreal, QC, Canada: Association for Computing Machinery, 2021. ISBN: 9781450384223. DOI: 10.1145/3472538.3472585. URL: <https://doi.org/10.1145/3472538.3472585>.
- [85] Khristopher J Brooks. "Twitter kills automatic photo-cropping feature after complaints of racial bias". In: *CBS News* (2021). URL: <https://www.cbsnews.com/news/twitter-kills-its-automatic-cropping-feature-after-complaints/>.
- [86] Khristopher J Brooks. "Twitter users say the platform crops out Black faces". In: *CBS News* (2020). URL: <https://www.cbsnews.com/news/twitter-image-cropping-algorithm-racial-profiling/>.
- [87] Andrew D Brown, Patrick Stacey, and Joe Nandhakumar. "Making sense of sensemaking narratives". In: *Human relations* 61.8 (2008), pp. 1035–1062.
- [88] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. "Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-Making in Child Welfare Services". In: (2019), 1–12. URL: <https://doi.org/10.1145/3290605.3300271>.
- [89] Finn Brunton and Helen Nissenbaum. *Obfuscation: A User's Guide for Privacy and Protest*. The MIT Press, 2015. ISBN: 0262029731.
- [90] Joy Buolamwini. "How I'm fighting bias in algorithms". In: *November, TEDx Beacon Street*[Video File] (2016).
- [91] Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. PMLR, 2018, pp. 77–91. URL: <https://proceedings.mlr.press/v81/buolamwini18a.html>.
- [92] Johanna Burai. *WORLD WHITE WEB*. 2015. URL: <https://johannaburai.com/World-White-Web>.

- [93] U.S. Census Bureau. In: *United States Census 2020* (2020). URL: <https://www2.census.gov/programs-surveys/decennial/2020/technical-documentation/questionnaires-and-instructions/questionnaires/2020-informational-questionnaire.pdf>.
- [94] Matt Burgess, Evaline Schot, and Gabriel Geiger. *This Algorithm Could Ruin Your Life*. 2023. URL: <https://www.wired.com/story/welfare-algorithms-discrimination/>.
- [95] Jenna Burrell. “Automated decision-making as domination”. In: *First Monday* 29.4 (2024). DOI: 10.5210/fm.v29i4.13630. URL: <https://firstmonday.org/ojs/index.php/fm/article/view/13630>.
- [96] Stephen Bush. *Beware the rise of the black box algorithm*. 2022. URL: <https://www.ft.com/content/3d5556c5-520e-497a-aa5e-2546c5bc50cf>.
- [97] Max C. *Unsell Yourself - A Protest Model Against Facebook*. 2011. URL: <https://yalelawtech.org/2011/05/10/unsell-yourself-a-protest-model-against-facebook/>.
- [98] Ángel Alexander Cabrera, Abraham J. Druck, Jason I. Hong, and Adam Perer. “Discovering and Validating AI Errors With Crowdsourced Failure Reports”. In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW2 (Oct. 2021). DOI: 10.1145/3479569. URL: <https://doi.org/10.1145/3479569>.
- [99] Dana Calacci and Alex Pentland. “Bargaining with the Black-Box: Designing and Deploying Worker-Centric Tools to Audit Algorithmic Management”. In: *Proc. ACM Hum.-Comput. Interact.* 6.CSCW2 (Nov. 2022). DOI: 10.1145/3570601. URL: <https://doi.org/10.1145/3570601>.
- [100] Chris Callison-Burch and Mark Dredze. “Creating speech and language data with amazon’s mechanical turk”. In: *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon’s Mechanical Turk*. 2010, pp. 1–12.
- [101] Michel Callon. “Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieuc Bay”. In: *The Sociological Review* 32.S1 (1984), pp. 196–233. DOI: <https://doi.org/10.1111/j.1467-954X.1984.tb00113.x>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-954X.1984.tb00113.x>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-954X.1984.tb00113.x>.
- [102] Steve Campbell, Melanie Greenwood, Sarah Prior, Toniele Shearer, Kerrie Walkem, Sarah Young, Danielle Bywaters, and Kim Walker. “Purposive sampling: complex or simple? Research case examples”. In: *Journal of research in Nursing* 25.8 (2020), pp. 652–661.
- [103] E. Carmi and P. Nakou. *What Mobilises People Against Big-Tech?* This report is published under a CC-BY license (Creative Commons Attribution 4.0 International License). London, UK, 2023. URL: <https://openaccess.city.ac.uk/id/eprint/31373/>.
- [104] Carlos del Castillo. *Multada una fábrica de Alicante por hacer reconocimiento facial a sus empleados sin avisarles*. 2023. URL: https://www.eldiario.es/tecnologia/multada-fabrica-alicante-reconocimiento-facial-empleados-avisarles_1_10152700.html.

- [105] Stephen Cave, Kanta Dihal, and Sarah Dillon. "Introduction: Imagining AI". In: *AI Narratives: A History of Imaginative Thinking about Intelligent Machines*. Oxford University Press, Feb. 2020. ISBN: 9780198846666. DOI: [10.1093/oso/9780198846666.003.0001](https://doi.org/10.1093/oso/9780198846666.003.0001). eprint: <https://academic.oup.com/book/0/chapter/321630469/chapter-pdf/57690562/oso-9780198846666-chapter-1.pdf>. URL: <https://doi.org/10.1093/oso/9780198846666.003.0001>.
- [106] DJ PATIL CECILIA MUÑOZ MEGAN SMITH. "Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights". In: (2016). URL: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.
- [107] Joel Chan, Steven Dang, and Steven P Dow. "Improving crowd innovation with expert facilitation". In: *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 2016, pp. 1223–1235.
- [108] Stevie Chancellor, Michael L. Birnbaum, Eric D. Caine, Vincent M. B. Silenzio, and Munmun De Choudhury. "A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Atlanta, GA, USA: Association for Computing Machinery, 2019, 79–88. ISBN: 9781450361255. DOI: [10.1145/3287560.3287587](https://doi.org/10.1145/3287560.3287587). URL: <https://doi.org/10.1145/3287560.3287587>.
- [109] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. "Fairness under unawareness: Assessing disparity when protected class is unobserved". In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 339–348.
- [110] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. "Investigating the Impact of Gender on Rank in Resume Search Engines". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2018, 1–14. ISBN: 9781450356206. URL: <https://doi.org/10.1145/3173574.3174225>.
- [111] Zhilong Chen, Hancheng Cao, Fengli Xu, Mengjie Cheng, Tao Wang, and Yong Li. "Understanding the Role of Intermediaries in Online Social E-commerce: An Exploratory Study of Beidian". In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2 (2020), pp. 1–24.
- [112] Hao-Fei Cheng, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghui Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. "How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions". 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: [10.1145/3491102.3501831](https://doi.org/10.1145/3491102.3501831). URL: <https://doi.org/10.1145/3491102.3501831>.
- [113] Myisha Cherry. *Anger Can Build a Better World*. 2020. URL: <https://www.theatlantic.com/ideas/archive/2020/08/how-anger-can-build-better-world/615625/>.
- [114] Rumman Chowdhury and Jutta Williams. "Introducing Twitter's first algorithmic bias bounty challenge". In: *Twitter* (2021). URL: https://blog.twitter.com/engineering/en_us/topics/insights/2021/algorithmic-bias-bounty-challenge.

- [115] Graham Christensen. @grhmc: “Ah no worries, Twitter doesn’t show the black person because the background color is lighter. Not a systemic racism thing.” with images of the white person prioritized regardless of background shade. 2020. URL: <https://twitter.com/grhmc/status/1307435994246008844>.
- [116] Graham Christensen. @grhmc: “I wonder how it is that you’ve said this without testing it?” with images testing multiple backgrounds in response to tweet claiming “...it’s the brighter background being used to make the cropping decision.” 2020. URL: <https://twitter.com/grhmc/status/1307435775856910343>.
- [117] Chris Cieslik. *We Didn’t Playtest This at All*. 2007. URL: <https://boardgamegeek.com/boardgame/31016/we-didnt-playtest-this-at-all>.
- [118] Marika Cifor, Patricia Garcia, T.L. Cowan, Jasmine Rault, Tonia Sutherland, Anita Say Chan, Jennifer Rode, Anna Lauren Hoffmann, Niloufar Salehi, Lisa Nakamura, and et al. *Feminist Data Manifest-No*. 2019. URL: <https://www.manifestno.com/>.
- [119] Victoria Clarke and Virginia Braun. “Thematic analysis”. In: *The Journal of Positive Psychology* 12.3 (2017), pp. 297–298. DOI: 10.1080/17439760.2016.1262613. eprint: <https://doi.org/10.1080/17439760.2016.1262613>. URL: <https://doi.org/10.1080/17439760.2016.1262613>.
- [120] Karen Clayton. *A company tracked its employees using facial recognition*. 2023. URL: <https://thenationview.com/world-news/180989.html>.
- [121] CLEAR. *CLEAR Lab Book: A living manual of our values, guidelines, and protocols*, V.03. St. John’s, NL, 2021. URL: <https://civiclaboratory.nl/clear-lab-book/>.
- [122] Will Conybeare and Rachel Menitoff. *Vandals, thieves attacking L.A. food delivery robots*. 2023. URL: <https://ktla.com/news/local-news/food-delivery-robots-under-attack-from-vandals-thieves-local-businesses-starting-to-be-affected/>.
- [123] Eric Corbett, Emily Denton, and Sheena Erete. “Power and Public Participation in AI”. In: *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. EAAMO ’23. Boston, MA, USA: Association for Computing Machinery, 2023. ISBN: 9798400703812. DOI: 10.1145/3617694.3623228. URL: <https://doi.org/10.1145/3617694.3623228>.
- [124] Susana Raquel Costa, Mirian Nogueira Tavares, Bruno Mendes Silva, Paulo Falcão Alves, Filipa Cerol, and Beatriz Isca. “Playing Against Hate Speech: How Teens See Hate Speech in Video Games and Online Gaming Communities”. In: *Journal of Digital Media & Interaction* 3.6 (2020), 34–52. DOI: 10.34624/jdmi.v3i6.15064. URL: <https://doi.org/10.34624/jdmi.v3i6.15064>.
- [125] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. “Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem”. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, 1571–1583. ISBN: 9781450393522. DOI: 10.1145/3531146.3533213. URL: <https://doi.org/10.1145/3531146.3533213>.
- [126] Rachel Courtland. *Bias detectives: The researchers striving to make algorithms fair*. 2018. URL: <https://www.nature.com/articles/d41586-018-05469-3>.

- [127] Henriette Cramer, Jean Garcia-Gathright, Aaron Springer, and Sravana Reddy. "Assessing and addressing algorithmic bias in practice". In: *Interactions* 25.6 (Oct. 2018), 58–63. ISSN: 1072-5520. DOI: [10.1145/3278156](https://doi.org/10.1145/3278156). URL: <https://doi.org/10.1145/3278156>.
- [128] Justin Cranshaw and Aniket Kittur. "The polymath project: lessons from a successful online collaboration in mathematics". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '11. Vancouver, BC, Canada: Association for Computing Machinery, 2011, 1865–1874. ISBN: 9781450302289. DOI: [10.1145/1978942.1979213](https://doi.org/10.1145/1978942.1979213). URL: <https://doi.org/10.1145/1978942.1979213>.
- [129] Kate Crawford. "The trouble with bias". In: *Conference on Neural Information Processing Systems, invited speaker*. 2017.
- [130] Kate Crawford and Trevor Paglen. "Excavating AI: The politics of images in machine learning training sets". In: *AI & Society* (2021), 1105–1116. DOI: [10.1007/s00146-021-01162-8](https://doi.org/10.1007/s00146-021-01162-8). URL: <https://doi.org/10.1007/s00146-021-01162-8>.
- [131] Michael Crider. *One rebel's malicious 'tar pit' trap is driving AI web-scrapers insane*. 2025. URL: <https://www.pcworld.com/article/2592071/one-rebels-malicious-tar-pit-trap-is-driving-ai-scrapers-insane.html>.
- [132] Erica Principe Cruz. "Counterspace Game Elements for This Pansexual Pilipina American Player's Joy, Rest, and Healing: An Autoethnographic Case Study of Playing Stardew Valley". In: *gamevironments* 17 (2022), pp. 314–354.
- [133] Erica Principe Cruz, Nalyn Sriwattanakomen, Jessica Hammer, and Geoff Kaufman. "Counterspace Games for BIWOC STEM Students". In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI EA '21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380959. DOI: [10.1145/3411763.3451811](https://doi.org/10.1145/3411763.3451811). URL: <https://doi.org/10.1145/3411763.3451811>.
- [134] Laudelino D. *yelps filtering! seems like extortion!* 2016. URL: <https://www.yelp.com/topic/miami-yelps-filtering-seems-like-extortion>.
- [135] Samantha Dalal, Siobhan Mackenzie Hall, and Nari Johnson. *Provocation: Who benefits from "inclusion" in Generative AI?* 2024. arXiv: [2411.09102](https://arxiv.org/abs/2411.09102) [cs.CY]. URL: <https://arxiv.org/abs/2411.09102>.
- [136] Jeffrey Dastin. "Amazon scraps secret AI recruiting tool that showed bias against women". In: (2018). URL: <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>.
- [137] Kevan Davis. *Dvorak*. 2000. URL: https://dvorakgame.co.uk/index.php/Main_Page.
- [138] Richard Dawkins. *The selfish gene*. Oxford University Press, 2016.
- [139] Michel De Certeau. *The practice of everyday life*, trans Steven Rendall. Berkeley: University of California Press, 1984.
- [140] Mack DeGeurin. *Welp, there goes twitter's ethical AI team, among others as employees post final messages*. 2022. URL: <https://gizmodo.com/twitter-layoffs-elon-musk-ai-ethics-1849743051>.

- [141] Wesley Hanwen Deng, Boyuan Guo, Alicia DeVrio, Hong Shen, Motahhare Eslami, and Kenneth Holstein. "Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: [10.1145/3544548.3581026](https://doi.org/10.1145/3544548.3581026). URL: <https://doi.org/10.1145/3544548.3581026>.
- [142] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. "Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits". en. In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. Seoul Republic of Korea: ACM, June 2022, pp. 473–484. ISBN: 978-1-4503-9352-2. DOI: [10.1145/3531146.3533113](https://dl.acm.org/doi/10.1145/3531146.3533113). URL: <https://dl.acm.org/doi/10.1145/3531146.3533113> (visited on 08/10/2022).
- [143] Melissa Densmore. "Claim Mobile: When to Fail a Technology". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '12. Austin, Texas, USA: Association for Computing Machinery, 2012, 1833–1842. ISBN: 9781450310154. DOI: [10.1145/2207676.2208319](https://doi.org/10.1145/2207676.2208319). URL: <https://doi.org/10.1145/2207676.2208319>.
- [144] dev. @kophing_: "DEAR TWITTER ARTISTS: SUFFER NO MORE WITH WEIRD CROPPING, HERE ARE TWITTERS OFFICIAL IMAGE RATIOS!" with images showing crop ratios depending on the number of photos included. 2018. URL: https://twitter.com/kophing_/status/1028000217654652928.
- [145] Michael A. DeVito, Jeremy Birnholtz, Jeffery T. Hancock, Megan French, and Sunny Liu. "How People Form Folk Theories of Social Media Feeds and What It Means for How We Study Self-Presentation". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2018, 1–12. ISBN: 9781450356206. URL: <https://doi.org/10.1145/3173574.3173694>.
- [146] 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*. CHI '17. Denver, Colorado, USA: Association for Computing Machinery, 2017, 3163–3174. ISBN: 9781450346559. DOI: [10.1145/3025453.3025659](https://doi.org/10.1145/3025453.3025659). URL: <https://doi.org/10.1145/3025453.3025659>.
- [147] Michael Ann DeVito. "Adaptive Folk Theorization as a Path to Algorithmic Literacy on Changing Platforms". In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW2 (Oct. 2021). DOI: [10.1145/3476080](https://doi.org/10.1145/3476080). URL: <https://doi.org/10.1145/3476080>.
- [148] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. "Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior". 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: [10.1145/3491102.3517441](https://doi.org/10.1145/3491102.3517441). URL: <https://doi.org/10.1145/3491102.3517441>.

- [149] Alicia DeVrio, Motahhare Eslami, and Kenneth Holstein. “Building, Shifting, & Employing Power: A Taxonomy of Responses From Below to Algorithmic Harm”. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’24. Rio de Janeiro, Brazil: Association for Computing Machinery, 2024, 1093–1106. ISBN: 9798400704505. DOI: [10.1145/3630106.3658958](https://doi.org/10.1145/3630106.3658958). URL: <https://doi.org/10.1145/3630106.3658958>.
- [150] Alicia DeVrio and David Gray Widder. “No More Plaudits for Algorithm Audits: A Provocation”. In: *Proceedings of the 2026 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’26. Malmö, Sweden: Association for Computing Machinery, 2026.
- [151] Nicholas Diakopoulos. “Algorithmic accountability reporting: On the investigation of black boxes”. In: *The Tow Center for Digital Journalism* (2014).
- [152] Nathalie Diberardino, Clair Baleshta, and Luke Stark. “Algorithmic Harms and Algorithmic Wrongs”. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’24. Rio de Janeiro, Brazil: Association for Computing Machinery, 2024, 1725–1732. ISBN: 9798400704505. DOI: [10.1145/3630106.3659001](https://doi.org/10.1145/3630106.3659001). URL: <https://doi.org/10.1145/3630106.3659001>.
- [153] William Dieterich, Christina Mendoza, and Tim Brennan. *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*. 2016. URL: https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.
- [154] Anna Dobrosovestnova, David J. Bailey, Ralf Vetter, and Masoumeh Mansouri. “Robot Vandalism: A Senseless Act?” In: *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction*. HRI ’26. Edinburgh, Scotland, UK: Association for Computing Machinery, 2026, 1322–1331. ISBN: 9798400721281. DOI: [10.1145/3757279.3788813](https://doi.org/10.1145/3757279.3788813). URL: <https://doi.org/10.1145/3757279.3788813>.
- [155] Nathan Doctor, Guy Fiennes, and Ciarán O’Connor. *NazTok: An organized neo-Nazi TikTok network is getting millions of views*. 2024. URL: https://www.isdglobal.org/digital_dispatches/naztok-an-organized-neo-nazi-tiktok-network-is-getting-millions-of-views/.
- [156] Kutsal Dogan. @dikili: “Bu yeni bir şey değil ama bir kere daha deneyelim. Twitter foto algoritması bakalım ne yapacak. Şaşırtmadı beni.” with related images. 2020. URL: <https://twitter.com/dikili/status/1326674332135854080>.
- [157] Bryan Dosono and Bryan Semaan. “Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on Reddit”. In: *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019, pp. 1–13.
- [158] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. “Shepherding the crowd yields better work”. In: *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. 2012, pp. 1013–1022.
- [159] Peter Drahos. “Information feudalism in the information society”. In: *The Information Society - TIS* 11 (July 1995), pp. 209–222. DOI: [10.1080/01972243.1995.9960193](https://doi.org/10.1080/01972243.1995.9960193).
- [160] Veena Dubal. “ON ALGORITHMIC WAGE DISCRIMINATION”. In: *Columbia Law Review* 123.7 (2023), pp. 1929–1992. ISSN: 00101958, 19452268. URL: <https://www.jstor.org/stable/27264954> (visited on 04/29/2024).

- [161] Egg, Inc. 2016. URL: <https://apps.apple.com/us/app/egg-inc/id993492744>.
- [162] Upol Ehsan, Ranjit Singh, Jacob Metcalf, and Mark Riedl. "The Algorithmic Imprint". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, 1305–1317. ISBN: 9781450393522. DOI: [10.1145/3531146.3533186](https://doi.org/10.1145/3531146.3533186). URL: <https://doi.org/10.1145/3531146.3533186>.
- [163] Nima Elmi. *Is Big Tech Setting Africa Back?* 2020. URL: <https://foreignpolicy.com/2020/11/11/is-big-tech-setting-africa-back/>.
- [164] Sheena Erete, Yolanda Rankin, and Jakita Thomas. "A Method to the Madness: Applying an Intersectional Analysis of Structural Oppression and Power in HCI and Design". In: *ACM Trans. Comput.-Hum. Interact.* 30.2 (2023). ISSN: 1073-0516. DOI: [10.1145/3507695](https://doi.org/10.1145/3507695). URL: <https://doi.org/10.1145/3507695>.
- [165] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. "First I "like" It, Then I Hide It: Folk Theories of Social Feeds". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI '16. San Jose, California, USA: Association for Computing Machinery, 2016, 2371–2382. ISBN: 9781450333627. DOI: [10.1145/2858036.2858494](https://doi.org/10.1145/2858036.2858494). URL: <https://doi.org/10.1145/2858036.2858494>.
- [166] Motahhare Eslami, Sneha R. Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. "Communicating Algorithmic Process in Online Behavioral Advertising". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2018, 1–13. ISBN: 9781450356206. URL: <https://doi.org/10.1145/3173574.3174006>.
- [167] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. "'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*. CHI '15. Seoul, Republic of Korea: Association for Computing Machinery, 2015, 153–162. ISBN: 9781450331456. DOI: [10.1145/2702123.2702556](https://doi.org/10.1145/2702123.2702556). URL: <https://doi.org/10.1145/2702123.2702556>.
- [168] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. "'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* 11.1 (2017), pp. 62–71. DOI: [10.1609/icwsm.v11i1.14898](https://doi.org/10.1609/icwsm.v11i1.14898). URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/14898>.
- [169] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. "User Attitudes towards Algorithmic Opacity and Transparency in Online Reviewing Platforms". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland UK: Association for Computing Machinery, 2019, 1–14. ISBN: 9781450359702. DOI: [10.1145/3290605.3300724](https://doi.org/10.1145/3290605.3300724). URL: <https://doi.org/10.1145/3290605.3300724>.
- [170] Virginia Eubanks. *A Child Abuse Prediction Model Fails Poor Families*. 2018. URL: <https://www.wired.com/story/excerpt-from-automating-inequality/>.

- [171] Virginia Eubanks. *A response to Allegheny County DHS*. 2018. URL: <https://virginia-eubanks.com/2018/02/16/a-response-to-allegheny-county-dhs/>.
- [172] Virginia Eubanks. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, 2018. ISBN: 9781250074317.
- [173] John H. Evans. "Between Technocracy and Democratic Legitimation: A Proposed Compromise Position for Common Morality Public Bioethics". In: *Journal of Medicine and Philosophy* 31.3 (2006). PMID: 16760101, pp. 213–234. DOI: 10.1080/03605310600732834. eprint: <https://www.tandfonline.com/doi/pdf/10.1080/03605310600732834>. URL: <https://www.tandfonline.com/doi/abs/10.1080/03605310600732834>.
- [174] Shelley Evenson. "Directed storytelling: Interpreting experience for design". In: *Design Studies: Theory and research in graphic design* (2006), pp. 231–240.
- [175] Chris F. *Why does my Yelp review not show up on the company site?* 2016. URL: <https://www.yelp.com/topic/walden-why-does-my-yelp-review-not-show-up-on-the-company-site>.
- [176] Megan Farokhmanesh. *YouTube is still restricting and demonetizing LGBT videos - and adding anti-LGBT ads to some*. 2018. URL: <https://www.theverge.com/2018/6/4/17424472/youtube-lgbt-demonetization-ads-algorithm>.
- [177] Rosta Farzan. @frostaff: "To those who think data is neutral: Farsi does not have gendered pronoun, see the bias the in Google Translate! #DataBias #CodedBias" with a related image. 2021. URL: <https://twitter.com/frostaff/status/1376699332917870592>.
- [178] Michael Feffer, Anusha Sinha, Wesley H. Deng, Zachary C. Lipton, and Hoda Heidari. "Red-Teaming for Generative AI: Silver Bullet or Security Theater?" In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7.1 (2024), 421–437. DOI: 10.1609/aies.v7i1.31647. URL: <https://ojs.aaai.org/index.php/AIES/article/view/31647>.
- [179] Kristie Fisher, Scott Counts, and Aniket Kittur. "Distributed sensemaking: Improving sensemaking by leveraging the efforts of previous users". In: *Proceedings of the 2012 CHI Conference on Human Factors in Computing Systems*. 2012, pp. 247–256.
- [180] Rich Fitzgerald. *DHS Response to 'Automated Inequality' by Virginia Eubanks*. 2018. URL: <https://www.alleghenycounty.us/files/assets/county/v/1/services/dhs/documents/alleggheny-family-screening-tool/dhs-reply-to-eubanks.pdf>.
- [181] Laura Forlano and Anijo Mathew and. "From Design Fiction to Design Friction: Speculative and Participatory Design of Values-Embedded Urban Technology". In: *Journal of Urban Technology* 21.4 (2014), pp. 7–24. DOI: 10.1080/10630732.2014.971525. eprint: <https://doi.org/10.1080/10630732.2014.971525>. URL: <https://doi.org/10.1080/10630732.2014.971525>.
- [182] Diana E. Forsythe. *Studying Those Who Study Us: An Anthropologist in the World of Artificial Intelligence*. Ed. by David J. Hess. Stanford, CA: Stanford University Press, 2001. ISBN: 0804742030.
- [183] African Tax Administration Forum. *ATAF PUBLISHES AN APPROACH TO TAXING THE DIGITAL ECONOMY*. 2020. URL: <https://www.ataftax.org/ataf-publishes-an-approach-to-taxing-the-digital-economy>.

- [184] Geoffrey A. Fowler. "Judge dismisses suit against Yelp". In: *The Wall Street Journal* (Oct. 2011).
- [185] Sarah E. Fox, Samantha Shorey, Esther Y. Kang, Dominique Montiel Valle, and Estefania Rodriguez. "Patchwork: The Hidden, Human Labor of AI Integration within Essential Work". In: *Proc. ACM Hum.-Comput. Interact.* 7.CSCW1 (Apr. 2023). DOI: 10.1145/3579514. URL: <https://doi.org/10.1145/3579514>.
- [186] Nancy Fraser. "Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy". In: *Social Text* 25/26 (1990), pp. 56–80. ISSN: 01642472, 15271951. URL: <http://www.jstor.org/stable/466240> (visited on 02/05/2023).
- [187] Sara de Freitas. "Are Games Effective Learning Tools? A Review of Educational Games". In: *Journal of Educational Technology & Society* 21.2 (2018), pp. 74–84. ISSN: 11763647, 14364522. URL: <http://www.jstor.org/stable/26388380> (visited on 09/12/2024).
- [188] Sorelle Friedler. *Obfuscating Data to Prevent Discrimination*. 2017. URL: <https://www.obfuscationworkshop.org/2017/10/obfuscating-data-to-prevent-discrimination/>.
- [189] Sorelle Friedler, Ranjit Singh, Borhane Bili-Hamelin, Jacob Metcalf, and Brian J. Chen. *AI Red-Teaming Is Not a One-Stop Solution to AI Harms: Recommendations for Using Red-Teaming for AI Accountability*. 2023. URL: <https://datasociety.net/library/ai-red-teaming-is-not-a-one-stop-solution-to-ai-harms-recommendations-for-using-red-teaming-for-ai-accountability/>.
- [190] Batya Friedman and Helen Nissenbaum. "Bias in Computer Systems". In: *ACM Trans. Inf. Syst.* 14.3 (1996), 330–347. ISSN: 1046-8188. DOI: 10.1145/230538.230561. URL: <https://doi.org/10.1145/230538.230561>.
- [191] Brian Fung. "The Internet has unearthed more racist Google Maps results". In: *The Washington Post* (May 2015). URL: <https://www.washingtonpost.com/news/the-switch/wp/2015/05/21/the-internet-has-unearthed-more-racist-google-maps-results>.
- [192] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. "'I Wouldn't Say Offensive but...': Disability-Centered Perspectives on Large Language Models". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '23. Chicago, IL, USA: Association for Computing Machinery, 2023, 205–216. ISBN: 9798400701924. DOI: 10.1145/3593013.3593989. URL: <https://doi.org/10.1145/3593013.3593989>.
- [193] Maya Indira Ganesh and Emanuel Moss. "Resistance and refusal to algorithmic harms: Varieties of 'knowledge projects'". In: *Media International Australia* 183.1 (2022), pp. 90–106. DOI: 10.1177/1329878X221076288. eprint: <https://doi.org/10.1177/1329878X221076288>. URL: <https://doi.org/10.1177/1329878X221076288>.
- [194] Ben Gansky and Sean McDonald. "CounterFAccTual: How FAccT Undermines Its Organizing Principles". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, 1982–1992. ISBN: 9781450393522. DOI: 10.1145/3531146.3533241. URL: <https://doi.org/10.1145/3531146.3533241>.

- [195] Patricia Garcia, Tonia Sutherland, Niloufar Salehi, Marika Cifor, and Anubha Singh. "No! Re-Imagining Data Practices Through the Lens of Critical Refusal". In: *Proc. ACM Hum.-Comput. Interact.* 6.CSCW2 (2022). DOI: [10.1145/3557997](https://doi.org/10.1145/3557997). URL: <https://doi.org/10.1145/3557997>.
- [196] Timnit Gebru. @timnitGebru: "You know I noticed this exact thing but I didn't say anything. I was like maybe I'm just noticing every little thing now. Its nice to see men like @mraginsky calling it out. Thank you for being an ally." 2019. URL: <https://twitter.com/timnitGebru/status/1080697598170808321>.
- [197] James Paul Gee. "Learning and Games". In: *The Ecology of Games: Connecting Youth, Games, and Learning*. The John D., Catherine T. MacArthur Foundation Series on Digital Media, and Learning, 2008, pp. 21–40. DOI: [10.1162/dmal.9780262693646.021](https://doi.org/10.1162/dmal.9780262693646.021). eprint: <http://www.mitpressjournals.org/doi/pdf/10.1162/dmal.9780262693646.021>. URL: <http://www.mitpressjournals.org/doi/abs/10.1162/dmal.9780262693646.021>.
- [198] Steve Gehlbach. *Empty Waymos invade Atlanta neighborhood, circle cul-de-sac for hours with no passengers*. 2026. URL: <https://www.wsbtv.com/news/local/atlanta/empty-waymos-invade-atlanta-neighborhood-circle-cul-de-sac-hours-with-no-passengers/CSNV2G5CZFHHFP6BOH6YF5RCFU/>.
- [199] R Stuart Geiger. "Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space". In: *Information, Communication & Society* 19.6 (2016), pp. 787–803.
- [200] R Stuart Geiger and Aaron Halfaker. "Using edit sessions to measure participation in Wikipedia". In: *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. 2013, pp. 861–870.
- [201] R. Stuart Geiger, Udayan Tandon, Anoolia Gakhokidze, Lian Song, and Lilly Irani. "Making Algorithms Public: Reimagining Auditing from Matters of Fact to Matters of Concern". In: *International Journal of Communication* 18 (2024), pp. 634–655. URL: <https://ojs3.ijoc.org/index.php/ijoc/article/view/20811>.
- [202] Doug Geisler. *Beat the Boss Expanded Edition*. URL: <https://www.practicalfox.com/product/beat-the-boss-2nd-edition/>.
- [203] "Gender and Jobs in Online Image Searches". In: (2018). URL: <https://www.pewresearch.org/social-trends/2018/12/17/gender-and-jobs-in-online-image-searches/>.
- [204] Marissa Kumar Gerchick, Ro Encarnación, Cole Tanigawa-Lau, Lena Armstrong, Ana Gutiérrez, and Danaé Metaxa. "Auditing the Audits: Lessons for Algorithmic Accountability from Local Law 144's Bias Audits". In: *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '25. New York, NY, USA: Association for Computing Machinery, 2025, 29–44. ISBN: 9798400714825. DOI: [10.1145/3715275.3732004](https://doi.org/10.1145/3715275.3732004). URL: <https://doi.org/10.1145/3715275.3732004>.
- [205] Martin Giles. *It's time to rein in the data barons*. 2018. URL: <https://www.technologyreview.com/2018/06/19/240453/its-time-to-rein-in-the-data-barons/>.
- [206] Tarleton Gillespie. "The relevance of algorithms". In: *Media technologies: Essays on communication, materiality, and society* 167.2014 (2014), p. 167.

- [207] Bruce Gilley. "Technocracy and democracy as spheres of justice in public policy". In: *Policy Sciences* 50.1 (Mar. 2017), pp. 9–22. ISSN: 1573-0891. DOI: [10.1007/s11077-016-9260-2](https://doi.org/10.1007/s11077-016-9260-2). URL: <https://doi.org/10.1007/s11077-016-9260-2>.
- [208] Bruce Glassco, Rob Daviau, Bill McQuillan, Mike Selinker, and Teeuwynn Woodruff. *Betrayal at House on the Hill*. 2004. URL: <https://boardgamegeek.com/boardgame/10547/betrayal-at-house-on-the-hill>.
- [209] Eric Goldman. "Court says Yelp doesn't extort businesses". In: *Forbes* (Sept. 2014). URL: <https://www.forbes.com/sites/ericgoldman/2014/09/03/court-says-yelp-doesnt-extort-businesses>.
- [210] Robert L. Goldstone, Sam Day, and Ji Y. Son. "Comparison". In: *Towards a theory of thinking*. Ed. by Britt Glatzeder and Albrecht Goel Vinod and Müller. Berlin, Heidelberg: Springer, 2010, pp. 103–121.
- [211] Hila Gonen and Yoav Goldberg. "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 609–614. DOI: [10.18653/v1/N19-1061](https://doi.org/10.18653/v1/N19-1061). URL: <https://aclanthology.org/N19-1061/>.
- [212] Bryce Goodman and Seth Flaxman. "EU regulations on algorithmic decision-making and a "right to explanation"". In: *ICML Workshop on Human Interpretability in Machine Learning*. 2016.
- [213] Ellen P. Goodman and Julia Trehu. *AI Audit Washing and Accountability*. 2022. URL: <https://ssrn.com/abstract=4227350>.
- [214] Rachel Goodman. *The Good Wife Tackles Algorithmic Discrimination. Meanwhile, in Real Life...* 2015. URL: <https://www.aclu.org/news/racial-justice/good-wife-tackles-algorithmic-discrimination-meanwhile>.
- [215] Google. *Empowering kids to be safe, confident explorers of the online world*. URL: <https://beinternetawesome.withgoogle.com/>.
- [216] "Google image search cements national stereotypes of 'racy' women". In: (). URL: <https://www.dw.com/en/google-image-search-cements-national-stereotypes-of-racy-women/a-56767605>.
- [217] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. "Jury learning: Integrating dissenting voices into machine learning models". In: *CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–19.
- [218] Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. "The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2021. ISBN: 9781450380966. URL: <https://doi.org/10.1145/3411764.3445423>.
- [219] André Gorz. "Strategy for Labor". In: *Theories of the Labor Movement*. Ed. by Bruce Nissen and Simeon Larson. Wayne State University Press, 1987, pp. 100–116.

- [220] Jeffrey Gottfried and Olivia Sidoti. *Teens and Video Games Today*. 2024. URL: <https://www.pewresearch.org/internet/2024/05/09/teens-and-video-games-today>.
- [221] Ben Green. "Data Science as Political Action: Grounding Data Science in a Politics of Justice". In: *Journal of Social Computing* 2.3 (2021), pp. 249–265. DOI: 10.23919/JSC.2021.0029. URL: <https://www.sciopen.com/article/10.23919/JSC.2021.0029>.
- [222] Catherine Grevet and Eric Gilbert. "Piggyback prototyping: Using existing, large-scale social computing systems to prototype new ones". In: *Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems*. 2015, pp. 4047–4056.
- [223] Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. "Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction". In: *WWW '18* (2018), 903–912. DOI: 10.1145/3178876.3186138. URL: <https://doi.org/10.1145/3178876.3186138>.
- [224] Alex Groce, Todd Kulesza, Chaoqiang Zhang, Shalini Shamasunder, Margaret Burnett, Weng-Keen Wong, Simone Stumpf, Shubhomoy Das, Amber Shinsel, Forrest Bice, et al. "You are the only possible oracle: Effective test selection for end users of interactive machine learning systems". In: *IEEE Transactions on Software Engineering* 40.3 (2013), pp. 307–323.
- [225] Lara Groves, Jacob Metcalf, Alayna Kennedy, Briana Vecchione, and Andrew Strait. "Auditing Work: Exploring the New York City algorithmic bias audit regime". In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '24. Rio de Janeiro, Brazil: Association for Computing Machinery, 2024, 1107–1120. ISBN: 9798400704505. DOI: 10.1145/3630106.3658959. URL: <https://doi.org/10.1145/3630106.3658959>.
- [226] Rosanna E. Guadagno, Daniel M. Rempala, Shannon Murphy, and Bradley M. Okdie. "What makes a video go viral? An analysis of emotional contagion and Internet memes". In: *Computers in Human Behavior* 29.6 (2013), pp. 2312–2319. ISSN: 0747-5632. DOI: <https://doi.org/10.1016/j.chb.2013.04.016>. URL: <https://www.sciencedirect.com/science/article/pii/S0747563213001192>.
- [227] Ben Guarino. "Google faulted for racial bias in image search results for black teenagers". In: (). URL: <https://www.dw.com/en/google-image-search-cements-national-stereotypes-of-racy-women/a-56767605>.
- [228] Eileen Guo and Tate Ryan-Mosley. *Computer scientists designing the future can't agree on what privacy means*. 2023. URL: <https://www.technologyreview.com/2023/04/03/1070665/cmu-university-privacy-battle-smart-building-sensors-mites/>.
- [229] Lauren Kaori Gurley. *Amazon's AI Cameras Are Punishing Drivers for Mistakes They Didn't Make*. 2021. URL: <https://www.vice.com/en/article/88npjv/amazons-ai-cameras-are-punishing-drivers-for-mistakes-they-didnt-make>.
- [230] Sireesh Gururaja, Amanda Bertsch, Clara Na, David Widder, and Emma Strubell. "To Build Our Future, We Must Know Our Past: Contextualizing Paradigm Shifts in Natural Language Processing". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor,

- Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 13310–13325. DOI: [10.18653/v1/2023.emnlp-main.822](https://doi.org/10.18653/v1/2023.emnlp-main.822). URL: <https://aclanthology.org/2023.emnlp-main.822/>.
- [231] Jessica Guynn. “Google photos labeled black people ‘gorillas’”. In: *USA Today* (June 2015). URL: <https://www.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-as-gorillas/29567465>.
- [232] Jürgen Habermas. *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. MIT press, 1989.
- [233] Aaron Halfaker and R. Stuart Geiger. “ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia”. In: *Proc. ACM Hum.-Comput. Interact.* 4.CSCW2 (2020). DOI: [10.1145/3415219](https://doi.org/10.1145/3415219). URL: <https://doi.org/10.1145/3415219>.
- [234] Aaron Halfaker, Aniket Kittur, and John Riedl. “Don’t bite the newbies: How reverts affect the quantity and quality of Wikipedia work”. In: *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*. 2011, pp. 163–172.
- [235] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. “Gender Recognition or Gender Reductionism? The Social Implications of Embedded Gender Recognition Systems”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI ’18. Montreal QC, Canada: Association for Computing Machinery, 2018, 1–13. ISBN: 9781450356206. DOI: [10.1145/3173574.3173582](https://doi.org/10.1145/3173574.3173582). URL: <https://doi.org/10.1145/3173574.3173582>.
- [236] Amber M. Hamilton. *Silicon Valley Pretends That Algorithmic Bias Is Accidental. It’s Not*. 2021. URL: <https://slate.com/technology/2021/07/silicon-valley-algorithmic-bias-structural-racism.html>.
- [237] Jessica Hammer. “Playing Prejudice: The Impact of Game-Play on Attributions of Racial and Gender Bias”. PhD thesis. Columbia University, 2014. DOI: [10.7916/D8BP00ZK](https://doi.org/10.7916/D8BP00ZK). URL: <https://doi.org/10.7916/D8BP00ZK>.
- [238] Jessica Hammer, Whitney Beltrán, Jonathan Walton, and Moyra Turkington. “Power and Control in Role-Playing Games”. In: *Role-Playing Game Studies*. 1st ed. Routledge, 2018, 448–467.
- [239] Jessica Hammer, Moyra Turkington, and Nathan LeBlanc. “Rosenstrasse: Holocaust Education Through Role Play”. In: *International Academic Conference on Meaningful Play*. East Lansing, MI, USA, 2018, pp. 1–28. URL: <http://replayable.net/research/rosenstrasse/>.
- [240] Anikó Hannák, Piotr Sapiezynski, Arash Molavi Khaki, David Lazer, Alan Mislove, and Christo Wilson. “Measuring Personalization of Web Search”. In: *CoRR abs/1706.05011* (2017). arXiv: [1706.05011](https://arxiv.org/abs/1706.05011). URL: <http://arxiv.org/abs/1706.05011>.
- [241] Anikó Hannák, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. “Measuring Price Discrimination and Steering on E-Commerce Web Sites”. In: *IMC ’14* (2014), 305–318. DOI: [10.1145/2663716.2663744](https://doi.org/10.1145/2663716.2663744). URL: <https://doi.org/10.1145/2663716.2663744>.
- [242] Donna Haraway. “Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective”. In: *Feminist Studies* 14.3 (1988), pp. 575–599. ISSN: 00463663. URL: <http://www.jstor.org/stable/3178066> (visited on 07/18/2023).

- [243] Christina Harrington, Sheena Erete, and Anne Marie Piper. "Deconstructing Community-Based Collaborative Design: Towards More Equitable Participatory Design Engagements". In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). DOI: 10.1145/3359318. URL: <https://doi.org/10.1145/3359318>.
- [244] Jamie Harris. *Facebook forced to ban its AI after it 'revealed how to make napalm and made racist comments*. 2022. URL: <https://www.the-sun.com/tech/6729391/meta-withdraws-ai-galactica-controversy/>.
- [245] Paisley Harris. "Gatekeeping and Remaking: The Politics of Respectability in African American Women's History and Black Feminism". English. In: *Journal of Women's History* 15.1 (2003). Name - Harvard University Press; University of Houston; University of North Carolina Press; Copyright - Copyright Indiana University Press Spring 2003; People - Wolcott, Victoria W; Schechter, Patricia; White, Frances; Last updated - 2023-11-28; SubjectsTermNotLitGenreText - Wolcott, Victoria W; Schechter, Patricia; White, Frances; Detroit Michigan; United States-US, pp. 212-220,238. URL: <https://www.proquest.com/scholarly-journals/gatekeeping-remaking-politics-respectability/docview/203247915/se-2>.
- [246] Caroline Haskins. *Dozens of Cities Have Secretly Experimented With Predictive Policing Software*. 2019. URL: <https://www.vice.com/en/article/dozens-of-cities-have-secretly-experimented-with-predictive-policing-software/>.
- [247] Nabil Hassein. *Against Black Inclusion in Facial Recognition*. 2017. URL: <https://digitaltalkingdrum.com/2017/08/15/against-black-inclusion-in-facial-recognition/>.
- [248] Andrew J. Hawkins. *Waymo autonomous vehicle attacked by an 'erratic' pedestrian in Arizona*. 2022. URL: <https://www.theverge.com/2022/7/7/23197041/waymo-self-driving-car-pedestrian-attack-arizona>.
- [249] Alex Hern. "Twitter apologises for 'racist' image-cropping algorithm". In: *The Guardian* (Sept. 2020). URL: <https://www.theguardian.com/technology/2020/sep/21/twitter-apologises-for-racist-image-cropping-algorithm>.
- [250] Joe Hernandez. *A parents' lawsuit accuses Amazon of selling suicide kits to teenagers*. 2022. URL: <https://www.npr.org/2022/10/09/1127686507/amazon-suicide-teenagers-poison>.
- [251] Evelyn Brooks Higginbotham. *Righteous Discontent: The Women's Movement in the Black Baptist Church: 1880-1920*. Harvard University Press, 1993.
- [252] Kashmir Hill. *I Cut the 'Big Five' Tech Giants From My Life. It Was Hell*. 2019. URL: <https://gizmodo.com/i-cut-the-big-five-tech-giants-from-my-life-it-was-hel-1831304194>.
- [253] Patricia. Hill Collins. *Black feminist thought : knowledge, consciousness, and the politics of empowerment*. eng. 2nd edition. Perspectives on gender. New York: Routledge, 2000. ISBN: 1-135-96013-5.
- [254] Kenneth Holstein, Erik Harpstead, Rebecca Gulotta, and Jodi Forlizzi. "Replay Enactments: Exploring Possible Futures through Historical Data". In: *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. DIS '20. Eindhoven, Netherlands: Association for Computing Machinery, 2020, 1607-1618. ISBN: 9781450369749. DOI: 10.1145/3357236.3395427. URL: <https://doi.org/10.1145/3357236.3395427>.

- [255] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. "Designing for complementarity: Teacher and student needs for orchestration support in AI-enhanced classrooms". In: (2019), pp. 157–171.
- [256] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. "Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?" In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, 1–16. ISBN: 9781450359702. DOI: 10.1145/3290605.3300830. URL: <https://doi.org/10.1145/3290605.3300830>.
- [257] Holy \$%#@: *Religion & Conscious Language*. panel at ACES Annual Conference. 2019. URL: <https://aceseditors.org/conference/past-conferences/2019/sessions>.
- [258] Sarah Homewood. "Inaction as a Design Decision: Reflections on Not Designing Self-Tracking Tools for Menopause". In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI EA '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, 1–12. ISBN: 9781450359719. DOI: 10.1145/3290607.3310430. URL: <https://doi.org/10.1145/3290607.3310430>.
- [259] Alexis Hope, Catherine D'Ignazio, Josephine Hoy, Rebecca Michelson, Jennifer Roberts, Kate Krontiris, and Ethan Zuckerman. "Hackathons as Participatory Design: Iterating Feminist Utopias". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, 1–14. ISBN: 9781450359702. DOI: 10.1145/3290605.3300291. URL: <https://doi.org/10.1145/3290605.3300291>.
- [260] Mokter Hossain. "Users' motivation to participate in online crowdsourcing platforms". In: *2012 International Conference on Innovation Management and Technology Research*. IEEE. 2012, pp. 310–315.
- [261] White House. "Big Data: A Report on Algorithmic Systems". In: *Opportunity, and Civil Rights* (2016).
- [262] Lara Houston, Steven J. Jackson, Daniela K. Rosner, Syed Ishtiaque Ahmed, Meg Young, and Laewoo Kang. "Values in Repair". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI '16. San Jose, California, USA: Association for Computing Machinery, 2016, 1403–1414. ISBN: 9781450333627. DOI: 10.1145/2858036.2858470. URL: <https://doi.org/10.1145/2858036.2858470>.
- [263] Josh Howarth. *TikTok User Age, Gender, & Demographics* (2024). 2024. URL: <https://explodingtopics.com/blog/tiktok-demographics#tiktok-users-age>.
- [264] Daniel C. Howe and Helen Nissenbaum. "Engineering privacy and protest: A case study of AdNauseam". English (US). In: *CEUR Workshop Proceedings* 1873 (2017). Funding Information: This publication has been supported in part by grants from US NSF CNS/NetS 105833, US NSF SATC 1642553, and the Research Grants Council of Hong Kong, China (Project No. CityU 11669616); 3rd International Workshop on Privacy Engineering, IWPE 2017 ; Conference date: 25-05-2017, pp. 57–64. ISSN: 1613-0073.

- [265] Yen-Chia Hsu, Himanshu Verma, Andrea Mauri, Illah Nourbakhsh, Alessandro Bozzon, et al. "Empowering local communities using artificial intelligence". In: *Patterns* 3.3 (2022), p. 100449.
- [266] Linda Huber. "'A Bridge to Nowhere': A Healthcare Case Study for Non-Reformist Design". In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. CHI '25. New York, NY, USA: Association for Computing Machinery, 2025. ISBN: 9798400713941. DOI: [10.1145/3706598.3713729](https://doi.org/10.1145/3706598.3713729). URL: <https://doi.org/10.1145/3706598.3713729>.
- [267] Linda Huber, Pedro Reynolds-Cuellar, Alicia DeVrio, Jensine Raihan, Cella M Sum, Lynn Dombrowski, Justine Zhang, Christoph B Becker, Lilly Irani, P M Krafft, and Margaret Hughes. "From Tech Lash to Tech Fash: Strategic Reflections on a Decade of Collective Organizing in Computing". In: *Adjunct Proceedings of the Sixth Decennial Aarhus Conference: Computing X Crisis*. AAR Adjunct '25. New York, NY, USA: Association for Computing Machinery, 2025. ISBN: 9798400719684. DOI: [10.1145/3737609.3747097](https://doi.org/10.1145/3737609.3747097). URL: <https://doi.org/10.1145/3737609.3747097>.
- [268] "I saw a tweet saying 'Google unprofessional hairstyles for work'. I did. Then I checked the 'professional' ones". In: (2016). URL: <https://twitter.com/HereroRocher/status/717457819864272896>.
- [269] Diana Ionescu. *After Self-Driving Taxis Hinder Emergency Response, Governors Highway Safety Association and Waymo Issue Guide for First Responders*. 2025. URL: <https://www.planetizen.com/news/2025/12/136591-after-self-driving-taxis-hinder-emergency-response-governors-highway-safety>.
- [270] Lilly C. Irani and M. Six Silberman. "Stories We Tell About Labor: Turkopticon and the Trouble with 'Design'". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI '16. San Jose, California, USA: Association for Computing Machinery, 2016, 4573–4586. ISBN: 9781450333627. DOI: [10.1145/2858036.2858592](https://doi.org/10.1145/2858036.2858592). URL: <https://doi.org/10.1145/2858036.2858592>.
- [271] Lilly C. Irani and M. Six Silberman. "Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '13. Paris, France: Association for Computing Machinery, 2013, 611–620. ISBN: 9781450318990. DOI: [10.1145/2470654.2470742](https://doi.org/10.1145/2470654.2470742). URL: <https://doi.org/10.1145/2470654.2470742>.
- [272] Neema Iyer. *Choose Your Own Fake News*. URL: <https://chooseyourownfakenews.com>.
- [273] Steven J. Jackson and Laewoo Kang. "Breakdown, Obsolescence and Reuse: HCI and the Art of Repair". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '14. Toronto, Ontario, Canada: Association for Computing Machinery, 2014, 449–458. ISBN: 9781450324731. DOI: [10.1145/2556288.2557332](https://doi.org/10.1145/2556288.2557332). URL: <https://doi.org/10.1145/2556288.2557332>.
- [274] Maurice Jakesch, Zana Bućinca, Saleema Amershi, and Alexandra Olteanu. "How Different Groups Prioritize Ethical Values for Responsible AI". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, 310–323. ISBN: 9781450393522. DOI: [10.1145/3531146.3533097](https://doi.org/10.1145/3531146.3533097). URL: <https://doi.org/10.1145/3531146.3533097>.

- [275] Jakob Linnaa Jensen. "Digital Feudalism". en. In: *The Medieval Internet: Power, Politics and Participation in the Digital Age*. Emerald Publishing Limited, 2020, pp. 95–109. ISBN: 978-1-83909-413-2 978-1-83909-412-5. DOI: [10.1108/978-1-83909-412-520201008](https://doi.org/10.1108/978-1-83909-412-520201008). URL: <https://www.emerald.com/insight/content/doi/10.1108/978-1-83909-412-520201008/full/html> (visited on 02/04/2023).
- [276] Sut Jhally and Bill Livant. "Watching as Working: The Valorization of Audience Consciousness". In: *Journal of Communication* 36.3 (1986), pp. 124–143. URL: <http://pinguet.free.fr/jhallylivant.pdf>.
- [277] Jazette Johnson, Vitica Arnold, Anne Marie Piper, and Gillian R. Hayes. "'It's a lonely disease': Cultivating Online Spaces for Social Support among People Living with Dementia and Dementia Caregivers". In: *Proc. ACM Hum.-Comput. Interact.* 6.CSCW2 (2022). DOI: [10.1145/3555133](https://doi.org/10.1145/3555133). URL: <https://doi.org/10.1145/3555133>.
- [278] Khari Johnson. "AI predictions for 2019 from Yann LeCun, Hilary Mason, Andrew Ng, and Rumman Chowdhury". In: *VentureBeat* (2019). URL: <https://venturebeat.com/2019/01/02/ai-predictions-for-2019-from-yann-lecun-hilary-mason-andrew-ng-and-rumman-chowdhury/>.
- [279] Khari Johnson. "Apparent racial bias found in Twitter photo algorithm". In: *VentureBeat* (2020). URL: <https://venturebeat.com/2020/09/20/apparent-racial-bias-found-in-twitter-photo-algorithm/>.
- [280] Nari Johnson, Sanika Moharana, Christina Harrington, Nazanin Andalibi, Hoda Heidari, and Motahhare Eslami. "The Fall of an Algorithm: Characterizing the Dynamics Toward Abandonment". In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '24. Rio de Janeiro, Brazil: Association for Computing Machinery, 2024, 337–358. ISBN: 9798400704505. DOI: [10.1145/3630106.3658910](https://doi.org/10.1145/3630106.3658910). URL: <https://doi.org/10.1145/3630106.3658910>.
- [281] Samuel GB Johnson, Greeshma Rajeev-Kumar, and Frank C Keil. "Sense-making under ignorance". In: *Cognitive psychology* 89 (2016), pp. 39–70.
- [282] Jobs with Justice and TESA Collective. *STRIKE! The Game of Worker Rebellion*. URL: <https://store.tesacollective.com/products/strike-the-game-of-worker-rebellion>.
- [283] *K-pop fans take over #whitelivesmatter hashtag*. 2020. URL: <https://www.nbcnews.com/news/asian-america/k-pop-fans-take-over-whitelivesmatter-hashtag-n1223376>.
- [284] Pratyusha Kalluri. *Don't ask if artificial intelligence is good or fair, ask how it shifts power*. 2020. URL: <https://www.nature.com/articles/d41586-020-02003-2>.
- [285] Inkoo Kang. "Businesses: 'Yelp is the thug of the Internet'". In: *MuckRock* (2013). URL: <https://www.muckrock.com/news/archives/2013/jan/23/businesses-yelp-thug-of-the-internet..>
- [286] Inkoo Kang. "Read nearly 700 FTC complaints regarding Yelp". In: *MuckRock* (2013). URL: <https://www.muckrock.com/news/archives/2013/jan/23/businesses-yelp-thug-of-the-internet/>.

- [287] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. "“Because AI is 100% right and safe”: User Attitudes and Sources of AI Authority in India". 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: [10.1145/3491102.3517533](https://doi.org/10.1145/3491102.3517533). URL: <https://doi.org/10.1145/3491102.3517533>.
- [288] Nadia Karizat, Dan Delmonaco, Motahhare Eslami, and Nazanin Andalibi. "Algorithmic Folk Theories and Identity: How TikTok Users Co-Produce Knowledge of Identity and Engage in Algorithmic Resistance". In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW2 (2021). DOI: [10.1145/3476046](https://doi.org/10.1145/3476046). URL: <https://doi.org/10.1145/3476046>.
- [289] Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Daniella Raz, and P. M. Krafft. "Toward situated interventions for algorithmic equity: lessons from the field". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Barcelona, Spain: Association for Computing Machinery, 2020, 45–55. ISBN: 9781450369367. DOI: [10.1145/3351095.3372874](https://doi.org/10.1145/3351095.3372874). URL: <https://doi.org/10.1145/3351095.3372874>.
- [290] Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. "More than fun and money. worker motivation in crowdsourcing—a study on mechanical turk". In: (2011).
- [291] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. "Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, 1–14. ISBN: 9781450367080. DOI: [10.1145/3313831.3376219](https://doi.org/10.1145/3313831.3376219). URL: <https://doi.org/10.1145/3313831.3376219>.
- [292] Anna Kawakami, Amanda Coston, Hoda Heidari, Kenneth Holstein, and Haiyi Zhu. "Studying Up Public Sector AI: How Networks of Power Relations Shape Agency Decisions Around AI Design and Use". In: *Proc. ACM Hum.-Comput. Interact.* 8.CSCW2 (Nov. 2024). DOI: [10.1145/3686989](https://doi.org/10.1145/3686989). URL: <https://doi.org/10.1145/3686989>.
- [293] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. "Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support". 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: [10.1145/3491102.3517439](https://doi.org/10.1145/3491102.3517439). URL: <https://doi.org/10.1145/3491102.3517439>.
- [294] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. "Unequal Representation and Gender Stereotypes in Image Search Results for Occupations". In: (2015), 3819–3828. URL: <https://doi.org/10.1145/2702123.2702520>.
- [295] Kate Kaye. *The FTC's new enforcement weapon spells death for algorithms*. 2022. URL: <https://www.protocol.com/policy/ftc-algorithm-destroy-data-privacy>.

- [296] Katherine C. Kellogg, Melissa A. Valentine, and Angèle Christin. "Algorithms at Work: The New Contested Terrain of Control". In: *Academy of Management Annals* 14.1 (2020), pp. 366–410. DOI: [10.5465/annals.2018.0174](https://doi.org/10.5465/annals.2018.0174). eprint: <https://doi.org/10.5465/annals.2018.0174>. URL: <https://doi.org/10.5465/annals.2018.0174>.
- [297] Os Keyes. "The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition". In: *Proc. ACM Hum.-Comput. Interact.* 2.CSCW (Nov. 2018). DOI: [10.1145/3274357](https://doi.org/10.1145/3274357). URL: <https://doi.org/10.1145/3274357>.
- [298] Os Keyes and Jeanie Austin. "Feeling fixes: Mess and emotion in algorithmic audits". In: *Big Data & Society* 9.2 (2022), p. 20539517221113772. DOI: [10.1177/20539517221113772](https://doi.org/10.1177/20539517221113772). eprint: <https://doi.org/10.1177/20539517221113772>. URL: <https://doi.org/10.1177/20539517221113772>.
- [299] Os Keyes, Jevan Hutson, and Meredith Durbin. "A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry". In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI EA '19. Glasgow, Scotland UK: Association for Computing Machinery, 2019, 1–11. ISBN: 9781450359719. DOI: [10.1145/3290607.3310433](https://doi.org/10.1145/3290607.3310433). URL: <https://doi.org/10.1145/3290607.3310433>.
- [300] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ring-shia, et al. "Dynabench: Rethinking benchmarking in NLP". In: *arXiv preprint arXiv:2104.14337* (2021).
- [301] Minseon Kim, Hyomin Lee, Boqing Gong, Huishuai Zhang, and Sung Ju Hwang. *Automatic Jailbreaking of the Text-to-Image Generative AI Systems*. 2024. arXiv: [2405.16567](https://arxiv.org/abs/2405.16567) [cs.AI]. URL: <https://arxiv.org/abs/2405.16567>.
- [302] Sara Kingsley, Proteeti Sinha, Clara Wang, Motahhare Eslami, and Jason I. Hong. "'Give Everybody [...] a Little Bit More Equity': Content Creator Perspectives and Responses to the Algorithmic Demonetization of Content Associated with Disadvantaged Groups". In: *Proc. ACM Hum.-Comput. Interact.* 6.CSCW2 (2022). DOI: [10.1145/3555149](https://doi.org/10.1145/3555149). URL: <https://doi.org/10.1145/3555149>.
- [303] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. "The future of crowd work". In: *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. CSCW '13. San Antonio, Texas, USA: Association for Computing Machinery, 2013, 1301–1318. ISBN: 9781450313315. DOI: [10.1145/2441776.2441923](https://doi.org/10.1145/2441776.2441923). URL: <https://doi.org/10.1145/2441776.2441923>.
- [304] Eric Klopfer, Scot Osterweil, and Katie Salen. *Moving learning games forward*. an Education Arcade paper (research report). 2009. URL: <https://hal.science/hal-00593085>.
- [305] Colin Koopman. *The power thinker: Original, painstaking, sometimes frustrating and often dazzling. Foucault's work on power matters now more than ever*. Ed. by SamEditor Haselby. 2017. URL: <https://aeon.co/essays/why-foucaults-work-on-power-is-more-important-than-ever>.

- [306] P. M. Krafft, Meg Young, Michael Katell, Jennifer E. Lee, Shankar Narayan, Micah Epstein, Dharma Dailey, Bernease Herman, Aaron Tam, Vivian Guetler, Corinne Bintz, Daniella Raz, Pa Ousman Jobe, Franziska Putz, Brian Robick, and Bissan Barghouti. "An Action-Oriented AI Policy Toolkit for Technology Audits by Community Advocates and Activists". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 2021, 772–781. ISBN: 9781450383097. DOI: [10.1145/3442188.3445938](https://doi.org/10.1145/3442188.3445938). URL: <https://doi.org/10.1145/3442188.3445938>.
- [307] Anna Kramer. *Twitter's image cropping was biased, so it dumped the algorithm*. 2021. URL: <https://www.protocol.com/twitter-image-cropping-algorithm-biased>.
- [308] Michael W Kramer. "Sensemaking". In: *The international encyclopedia of organizational communication* (2017), pp. 1–10.
- [309] Robert E Kraut and Paul Resnick. "Encouraging contribution to online communities". In: *Building successful online communities: Evidence-based social design* (2011), pp. 21–76.
- [310] Joshua A Kroll, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, and Harlan Yu. "Accountable algorithms". In: *U. Pa. L. Rev.* 165 (2016), p. 633.
- [311] James Kuczmariski. *Reducing gender bias in Google Translate*. 2018. URL: <https://blog.google/products/translate/reducing-gender-bias-google-translate/>.
- [312] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P. Gummadi, and Karrie Karahalios. "Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media". In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '17. Portland, Oregon, USA: Association for Computing Machinery, 2017, 417–432. ISBN: 9781450343350. DOI: [10.1145/2998181.2998321](https://doi.org/10.1145/2998181.2998321). URL: <https://doi.org/10.1145/2998181.2998321>.
- [313] Bogdan Kulynych, Rebekah Overdorf, Carmela Troncoso, and Seda Gürses. "POTs: protective optimization technologies". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Barcelona, Spain: Association for Computing Machinery, 2020, 177–188. ISBN: 9781450369367. DOI: [10.1145/3351095.3372853](https://doi.org/10.1145/3351095.3372853). URL: <https://doi.org/10.1145/3351095.3372853>.
- [314] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. "Identifying Unknown Unknowns in the Open World: Representations and Policies for Guided Exploration". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI'17. San Francisco, California, USA: AAAI Press, 2017, 2124–2132.
- [315] Michelle S. Lam, Mitchell L. Gordon, Danaë Metaxa, Jeffrey T. Hancock, James A. Landay, and Michael S. Bernstein. "End-User Audits: A System Empowering Communities to Lead Large-Scale Investigations of Harmful Algorithmic Behavior". In: *Proc. ACM Hum.-Comput. Interact.* 6.CSCW2 (2022). DOI: [10.1145/3555625](https://doi.org/10.1145/3555625). URL: <https://doi.org/10.1145/3555625>.

- [316] Max Larson. "Computer Center Sabotage, 1968–1971: Luddism, Black Studies, and the Diversion of Technological Progress". In: *boundary* 2 50.4 (Nov. 2023), pp. 195–226. ISSN: 0190-3659. DOI: [10.1215/01903659-10694281](https://doi.org/10.1215/01903659-10694281). eprint: <https://read.dukeupress.edu/boundary-2/article-pdf/50/4/195/2031865/1951arson.pdf>. URL: <https://doi.org/10.1215/01903659-10694281>.
- [317] Max Larson. *Computer Protest & Sabotage*. 2020. URL: <https://loadedonloaded.github.io/sabotage/>.
- [318] Jonathan Lazar. "Public Policy and HCI: Making an Impact in the Future". In: *Interactions* 22.5 (2015), 69–71. ISSN: 1072-5520. DOI: [10.1145/2807916](https://doi.org/10.1145/2807916). URL: <https://doi.org/10.1145/2807916>.
- [319] Christopher A. Le Dantec. "Design through Collective Action / Collective Action through Design". In: *Interactions* 24.1 (2016), 24–30. ISSN: 1072-5520. DOI: [10.1145/3018005](https://doi.org/10.1145/3018005). URL: <https://doi.org/10.1145/3018005>.
- [320] Matt Leacock. *Pandemic*. 2008. URL: <https://www.zmangames.com/game/pandemic/>.
- [321] The Algorithmic Justice League. *The Algorithmic Justice League: Mission, Team and Story*. URL: <https://www.ajl.org/about>.
- [322] Theodore M. Lechterman. "The Concept of Accountability in AI Ethics and Governance". In: *The Oxford Handbook of AI Governance*. Oxford University Press, 2022. ISBN: 9780197579329. DOI: [10.1093/oxfordhb/9780197579329.013.10](https://doi.org/10.1093/oxfordhb/9780197579329.013.10). URL: <https://doi.org/10.1093/oxfordhb/9780197579329.013.10>.
- [323] Michelle Seng Ah Lee and Jat Singh. "The Landscape and Gaps in Open Source Fairness Toolkits". 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: [10.1145/3411764.3445261](https://doi.org/10.1145/3411764.3445261). URL: <https://doi.org/10.1145/3411764.3445261>.
- [324] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. "Meme-tracking and the dynamics of the news cycle". In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2009, 497–506. DOI: [10.1145/1557019.1557077](https://doi.org/10.1145/1557019.1557077). URL: <https://doi.org/10.1145/1557019.1557077>.
- [325] Hanlin Li, Nicholas Vincent, Stevie Chancellor, and Brent Hecht. "The Dimensions of Data Labor: A Road Map for Researchers, Activists, and Policymakers to Empower Data Producers". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '23. Chicago, IL, USA: Association for Computing Machinery, 2023, 1151–1161. ISBN: 9798400701924. DOI: [10.1145/3593013.3594070](https://doi.org/10.1145/3593013.3594070). URL: <https://doi.org/10.1145/3593013.3594070>.
- [326] Hanlin Li, Nicholas Vincent, Janice Tsai, Jofish Kaye, and Brent Hecht. "How Do People Change Their Technology Use in Protest? Understanding". In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (2019). DOI: [10.1145/3359189](https://doi.org/10.1145/3359189). URL: <https://doi.org/10.1145/3359189>.
- [327] Calvin A. Liang, Sean A. Munson, and Julie A. Kientz. "Embracing Four Tensions in Human-Computer Interaction Research with Marginalized People". In: *ACM Trans. Comput.-Hum. Interact.* 28.2 (Apr. 2021). ISSN: 1073-0516. DOI: [10.1145/3443686](https://doi.org/10.1145/3443686). URL: <https://doi.org/10.1145/3443686>.

- [328] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. “Blaming Humans and Machines: What Shapes People’s Reactions to Algorithmic Harm”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3580953. URL: <https://doi.org/10.1145/3544548.3580953>.
- [329] Kimberly Ling, Gerard Beenen, Pamela Ludford, Xiaoqing Wang, Klarissa Chang, Xin Li, Dan Cosley, Dan Frankowski, Loren Terveen, Al Mamunur Rashid, et al. “Using social psychology to motivate contributions to online communities”. In: *Journal of Computer-Mediated Communication* 10.4 (2005), pp. 00–00.
- [330] Jean Linis-Dinco. @imajeanepeace: “In the #Philippines, we don’t distinguish #gender through #pronouns. ‘Siya’ (He/She/Them) is gender-neutral. However, look at how #Google Translate gendered the pronouns according to the adjectives and the job. | Tried this after seeing the Hungarian-English bias. #CodedBias” with a related image. 2021. URL: <https://twitter.com/imajeanepeace/status/1374135940898156544>.
- [331] Rosina Lippi-Green. “English with an Accent: Language, Ideology and Discrimination in the United States”. In: Routledge, 2012. Chap. Language Subordination.
- [332] Anthony Liu, Santiago Guerra, Isaac Fung, Gabriel Matute, Ece Kamar, and Walter Lasecki. “Towards hybrid human-AI workflows for unknown unknown detection”. In: *Proceedings of The Web Conference 2020*. 2020, pp. 2432–2442.
- [333] John Locke. *Two Treatises of Government*. Ed. by Peter Laslett. The New American Library, 1960.
- [334] Shayne Longpre, Kevin Klyman, Ruth E. Appel, Sayash Kapoor, Rishi Bommasani, Michelle Sahar, Sean McGregor, Avijit Ghosh, Borhane Blili-Hamelin, Nathan Butters, Alondra Nelson, Amit Elazari, Andrew Sellars, Casey John Ellis, Dane Sherrets, Dawn Song, Harley Geiger, Ilona Cohen, Lauren McIlvenny, Madhulika Srikumar, Mark M. Jaycox, Markus Anderljung, Nadine Farid Johnson, Nicholas Carlini, Nicolas Miall, Nik Marda, Peter Henderson, Rebecca S. Portnoff, Rebecca Weiss, Victoria Westerhoff, Yacine Jernite, Rumman Chowdhury, Percy Liang, and Arvind Narayanan. *In-House Evaluation Is Not Enough: Towards Robust Third-Party Flaw Disclosure for General-Purpose AI*. 2025. arXiv: 2503.16861 [cs.AI]. URL: <https://arxiv.org/abs/2503.16861>.
- [335] Audre Lorde. *Sister Outsider: Essays and Speeches*. Trumansburg, NY: The Crossing Press, 1984, pp. 110–114. ISBN: 0895941422. URL: <https://www.proquest.com/books/sister-outsider-essays-speeches/docview/2138588223/section-2>.
- [336] Audre Lorde. *The Uses of Anger: Women responding to racism*. 1981. URL: <https://www.blackpast.org/african-american-history/speeches-african-american-history/1981-audre-lorde-uses-anger-women-responding-racism/>.
- [337] Kristine Lu. “Designing Democratic Systems for Civic Collective Action”. In: *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. CSCW ’21. Virtual Event, USA: Association for Computing Machinery, 2021, 270–274. ISBN: 9781450384797. DOI: 10.1145/3462204.3481792. URL: <https://doi.org/10.1145/3462204.3481792>.

- [338] Yiwen Lu. *Driverless Taxis Blocked Ambulance in Fatal Accident, San Francisco Fire Dept. Says*. 2023. URL: <https://www.nytimes.com/2023/09/02/technology/driverless-cars-cruise-san-francisco.html>.
- [339] Kristian Lum and Rumman Chowdhury. *What is an “algorithm”? it depends whom you ask*. 2021. URL: <https://www.technologyreview.com/2021/02/26/1020007/what-is-an-algorithm/>.
- [340] Kathleen Maas Weigert. “Structural Violence”. In: Dec. 2008, pp. 2004–2011. ISBN: 9780123739858. DOI: [10.1016/B978-012373985-8.00169-0](https://doi.org/10.1016/B978-012373985-8.00169-0).
- [341] Douglas MacMillan, David Ovalle, and Aaron Schaffer. *Arrested by AI: Police ignore standards after facial recognition matches*. 2025. URL: <https://www.washingtonpost.com/business/interactive/2025/police-artificial-intelligence-facial-recognition/>.
- [342] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. “Assessing the Fairness of AI Systems: AI Practitioners’ Processes, Challenges, and Needs for Support”. In: *arXiv preprint arXiv:2112.05675* (2021).
- [343] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. “Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery, 2020, 1–14. ISBN: 9781450367080. DOI: [10.1145/3313831.3376445](https://doi.org/10.1145/3313831.3376445). URL: <https://doi.org/10.1145/3313831.3376445>.
- [344] Colin Madland. @colinmadland: “A faculty member has been asking how to stop Zoom from removing his head when he uses a virtual background. We suggested the usual plain background, good lighting etc, but it didn’t work. I was in a meeting with him today when I realized why it was happening.” with corresponding images and concluding post: “Turns out @zoom_us has a crappy face-detection algorithm that erases black faces...and determines that a nice pale globe in the background must be a better face than what should be obvious.” 2020. URL: <https://twitter.com/colinmadland/status/1307111816250748933>.
- [345] Colin Madland. @colinmadland: “Flipped the image....@Twitter is trash.” with related image. 2020. URL: <https://twitter.com/colinmadland/status/1307130447671984129>.
- [346] Colin Madland. @colinmadland: “Geez...any guesses why @Twitter defaulted to show only the right side of the picture on mobile?” with related image. 2020. URL: <https://twitter.com/colinmadland/status/1307115534383710208>.
- [347] Suresh S Malladi and Hemang C Subramanian. “Bug bounty programs for cybersecurity: Practices, issues, and recommendations”. In: *IEEE Software* 37.1 (2019), pp. 31–39.
- [348] Rasmus Malver. @RasmusMalver: “Virker det med tegneseriepersoner?” with related images. 2020. URL: <https://twitter.com/RasmusMalver/status/1307615213810839552>.
- [349] Ellen B. Mandinach and Hugh F. Cline. *Classroom Dynamics: Implementing a Technology-Based Learning Environment*. Routledge, 2013. ISBN: 9781315044446. DOI: [10.4324/9781315044446](https://doi.org/10.4324/9781315044446). URL: <https://doi.org/10.4324/9781315044446>.
- [350] Mark. @MarkEMarkAU: “I tried it with dogs. Let’s see.” with related images. 2020. URL: <https://twitter.com/MarkEMarkAU/status/1307616892551487488>.

- [351] Bugcrowd Product Marketing. *Reducing Noise in Crowdsourced Security*. 2020. URL: <https://www.bugcrowd.com/blog/reducing-noise-in-crowdsourced-security/>.
- [352] Aarian Marshall. *Gig Workers Gather Their Own Data to Check the Algorithm's Math*. 2021. URL: <https://www.wired.com/story/gig-workers-gather-data-check-algorithm-math/>.
- [353] Mike Masnick, Randy Lubin, and Leigh Beadon. *Trust & Safety Tycoon*. 2023. URL: <https://trustandsafety.fun>.
- [354] Arunesh Mathur, Jessica Vitak, Arvind Narayanan, and Marshini Chetty. "Characterizing the Use of Browser-Based Blocking Extensions to Prevent Online Tracking". In: *Proceedings of the Fourteenth USENIX Conference on Usable Privacy and Security*. SOUPS '18. Baltimore, MD, USA: USENIX Association, 2018, 103–116. ISBN: 9781931971454.
- [355] Colleen McClain, Brian Kennedy, Jeffrey Gottfried, Monica Anderson, and Giancarlo Pasquini. *How the U.S. Public and AI Experts View Artificial Intelligence*. 2025. URL: <https://www.pewresearch.org/internet/2025/04/03/how-the-us-public-and-ai-experts-view-artificial-intelligence/>.
- [356] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. "Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice". In: *Proceedings of the ACM on human-computer interaction* 3.CSCW (2019), pp. 1–23.
- [357] Dean Richard McKinnel, Tooska Dargahi, Ali Dehghantanha, and Kim-Kwang Raymond Choo. "A systematic literature review and meta-analysis on artificial intelligence in penetration testing and vulnerability assessment". In: *Computers & Electrical Engineering* 75 (2019), pp. 175–188. ISSN: 0045-7906. DOI: <https://doi.org/10.1016/j.compeleceng.2019.02.022>. URL: <https://www.sciencedirect.com/science/article/pii/S0045790618315489>.
- [358] Morgan Meaker. *Algorithms Policed Welfare Systems For Years. Now They're Under Fire for Bias*. 2024. URL: <https://www.wired.com/story/algorithms-policed-welfare-systems-for-years-now-theyre-under-fire-for-bias/>.
- [359] Ivan Mehta. "Why Twitter's image cropping algorithm appears to have white bias". In: *TNW | Neural* (2021). URL: <https://thenextweb.com/news/why-twitters-image-cropping-algorithm-appears-to-have-white-bias>.
- [360] Brian Merchant. *Blood in the Machine: The origins of the rebellion against big tech*. Little, Brown and Company, 2023.
- [361] Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. "Auditing Algorithms: Understanding Algorithmic Systems from the Outside In". In: *Foundations and Trends® in Human-Computer Interaction* 14.4 (2021), pp. 272–344. ISSN: 1551-3955. DOI: [10.1561/11000000083](https://doi.org/10.1561/11000000083). URL: <http://dx.doi.org/10.1561/11000000083>.
- [362] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. "Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 2021, 735–746. ISBN: 9781450383097. DOI: [10.1145/3442188.3445935](https://doi.org/10.1145/3442188.3445935). URL: <https://doi.org/10.1145/3442188.3445935>.

- [363] Cade Metz. “‘Nerd,’ ‘Nonsmoker,’ ‘Wrongdoer’: How Might A.I. Label You?” In: *The New York Times* (2019). URL: <https://www.nytimes.com/2019/09/20/arts/design/imagenet-trevor-paglen-ai-facial-recognition.html>.
- [364] Milagros Miceli, Julian Posada, and Tianling Yang. “Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power?” In: *Proc. ACM Hum.-Comput. Interact.* 6.GROUP (2022). DOI: 10.1145/3492853. URL: <https://doi.org/10.1145/3492853>.
- [365] James Mills. *Driverless Vehicles Are the Newest Mass Surveillance Tool of Law Enforcement*. 2025. URL: <https://www.criminallegalnews.org/news/2025/jul/1/driverless-vehicles-are-newest-mass-surveillance-tool-law-enforcement/>.
- [366] Ronald B Mincy. “The Urban Institute audit studies: Their research and policy context”. In: *Clear and convincing evidence: Measurement of discrimination in America*. Ed. by M. Fix and R. J. Struyk. Urban Institute Press, 1993, pp. 165–86.
- [367] mitch. @nanobope: “TWITTER CROP GUIDE THREAD! (1/5)” with thread images showing the dimensions for crops sizes depending on how many are included. 2020. URL: <https://twitter.com/nanobop/status/1255002567131557888>.
- [368] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. “Model cards for model reporting”. In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 220–229.
- [369] Dwayne Monroe. *Manifesto on Algorithmic Sabotage: A Review*. 2024. URL: <https://monroelab.com/2024/04/02/manifesto-on-algorithmic-sabotage-a-review/>.
- [370] Dylan Moore, Rebecca Currano, Michael Shanks, and David Sirkin. “Defense Against the Dark Cars: Design Principles for Griefing of Autonomous Vehicles”. In: *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. HRI ’20. Cambridge, United Kingdom: Association for Computing Machinery, 2020, 201–209. ISBN: 9781450367462. DOI: 10.1145/3319502.3374796. URL: <https://doi.org/10.1145/3319502.3374796>.
- [371] m_paardekoper. @m_paardekoper: “Dark background”. 2020. URL: https://twitter.com/m_paardekoper/status/1307636655508140032.
- [372] m_paardekoper. @m_paardekoper: “McConnell 50% opacity to reduce contrast”. 2020. URL: https://twitter.com/m_paardekoper/status/1307641363756965891.
- [373] m_paardekoper. @m_paardekoper: “Obama 50% more contrast”. 2020. URL: https://twitter.com/m_paardekoper/status/1307636653683601408.
- [374] Michael Muller. “HCI as Translation Work: How Translation Studies can Inform HCI Research and Practice”. In: *Presentations at the CHI 2004 Workshop on Reflective HCI*. CHI ’04. New York, NY, USA: Association for Computing Machinery, 2004. URL: <https://dominoweb.draco.res.ibm.com/reports/rc23318.pdf>.
- [375] Martin Murer, Verena Fuchsberger, and Manfred Tscheligi. “Un-Crafting: De-Constructive Engagements with Interactive Artifacts”. In: *Proceedings of the Eleventh International Conference on Tangible, Embedded, and Embodied Interaction*. TEI ’17. Yokohama, Japan: Association for Computing Machinery, 2017, 67–77. ISBN: 9781450346764. DOI: 10.1145/3024969.3024993. URL: <https://doi.org/10.1145/3024969.3024993>.

- [376] Mahin Naderifar, Hamideh Goli, and Fereshteh Ghaljaie. "Snowball Sampling: A Purposeful Method of Sampling in Qualitative Research". In: *Strides in Development of Medical Education* 14.3 (2017), pp. –. ISSN: 2645-3525. DOI: 10.5812/sdme.67670. eprint: https://sdme.kmu.ac.ir/article_90598_3632edfb2e97c38d73c0bdea8753195c.pdf. URL: https://sdme.kmu.ac.ir/article_90598.html.
- [377] Dawn Nafus and Jamie Sherman. "This One Does Not Go Up To 11: The Quantified Self Movement as an Alternative Big Data Practice". In: *International Journal of Communication* 8.0 (2014). ISSN: 1932-8036. URL: <https://ijoc.org/index.php/ijoc/article/view/2170>.
- [378] Mengjun Ni, Jing Yang, Xin Lin, and Liang He. "Reducing unknown unknowns with guidance in image caption". In: *International Conference on Artificial Neural Networks*. Springer. 2017, pp. 547–555.
- [379] Safiya Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018. ISBN: 9781479849949. URL: <http://www.jstor.org/stable/j.ctt1pwt9w5> (visited on 09/09/2024).
- [380] Safiya Umoja Noble. "Google Search: Hyper-visibility as a Means of Rendering Black Women and Girls Invisible". In: *InVisible Culture: A Journal for Visual Culture* 19 (2013). URL: <http://hdl.handle.net/1802/28018>.
- [381] Safiya Umoja Noble. "Searching for black girls: old traditions in new media". PhD thesis. University of Illinois at Urbana-Champaign, 2012.
- [382] Besmira Nushi, Ece Kamar, and Eric Horvitz. "Towards accountable ai: Hybrid human-machine analyses for characterizing system failure". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 6. 2018, pp. 126–135.
- [383] Rodrigo Ochigame and Katherine Ye. "Search Atlas: Visualizing Divergent Search Results Across Geopolitical Borders". In: *Proceedings of the 2021 ACM Designing Interactive Systems Conference*. DIS '21. Virtual Event, USA: Association for Computing Machinery, 2021, 1970–1983. ISBN: 9781450384766. DOI: 10.1145/3461778.3462032. URL: <https://doi.org/10.1145/3461778.3462032>.
- [384] Adriaan Odendaal and Karla Zavala. *Black Boxes out of Cardboard: Algorithmic literacy through critical board game design*. 2018. URL: <https://analoggamestudies.org/2018/12/black-boxes-out-of-cardboard-algorithmic-literacy-through-critical-board-game-design>.
- [385] Victor Ojewale, Ryan Steed, Briana Vecchione, Abeba Birhane, and Inioluwa Deborah Raji. *Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling*. 2024. arXiv: 2402.17861 [cs.CY]. URL: <https://arxiv.org/abs/2402.17861>.
- [386] Chinasa T. Okolo, Srujana Kamath, Nicola Dell, and Aditya Vashistha. "'It cannot do all of my work': Community Health Worker Perceptions of AI-Enabled Mobile Health Applications in Rural India". 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: 10.1145/3411764.3445420. URL: <https://doi.org/10.1145/3411764.3445420>.

- [387] Parmy Olson. “The algorithm that helped google translate become sexist”. In: *Forbes* (Feb. 2018). URL: <https://www.forbes.com/sites/parmyolson/2018/02/15/the-algorithm-that-helped-google-translate-become-sexist>.
- [388] Alexandra Olteanu, Michael Ekstrand, Carlos Castillo, and Jina Suh. *Responsible AI Research Needs Impact Statements Too*. 2023. arXiv: 2311.11776 [cs.AI]. URL: <https://arxiv.org/abs/2311.11776>.
- [389] Jo Ann Oravec. “Rage against robots: Emotional and motivational dimensions of anti-robot attacks, robot sabotage, and robot bullying”. In: *Technological Forecasting and Social Change* 189 (2023), p. 122249. ISSN: 0040-1625. DOI: <https://doi.org/10.1016/j.techfore.2022.122249>. URL: <https://www.sciencedirect.com/science/article/pii/S0040162522007703>.
- [390] Darrell Owens. @IDoTheThinking: “Look a lotta people keep Q’ing the very first test to prove it’s not racist. The tests with natural lighting are more relevant to people. If a Black face is only chosen in studio light backgrounds then that means 99% of dark skinned faces in real life won’t be chosen.” with related images. 2020. URL: <https://twitter.com/IDoTheThinking/status/1307505161640660992>.
- [391] Darrell Owens. @IDoTheThinking: “Ok just one more test. A Black woman and a White woman are lying in the same bed together with the same lighting in this picture. What does Twitter’s Machine Learning focus on:” with related images. 2020. URL: <https://twitter.com/IDoTheThinking/status/1307449013247991808>.
- [392] Andy Ozment. “Bug auctions: Vulnerability markets reconsidered”. In: *Third Workshop on the Economics of Information Security*. 2004, pp. 19–26.
- [393] Mola Mola P. *Yelp filtering almost all reviews, removing more by the day*. 2018. URL: <https://www.yelp.com/topic/dallas-yelp-filtering-almost-all-reviews-removing-more-by-the-day>.
- [394] Pedro Pais, David Gonçalves, Daniel Reis, João Cadete Nunes Godinho, João Filipe Morais, Manuel Piçarra, Pedro Trindade, Dmitry Alexandrovsky, Kathrin Gerling, João Guerreiro, and André Rodrigues. “A Living Framework for Understanding Cooperative Games”. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. CHI ’24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: [10.1145/3613904.3641953](https://doi.org/10.1145/3613904.3641953). URL: <https://doi.org/10.1145/3613904.3641953>.
- [395] Francesca Palmiotto. *Challenging automated filtering systems - The case of Yelp*. 2020. URL: <https://ai-laws.org/en/2020/05/challenging-automated-filtering-systems-the-case-of-yelp/>.
- [396] Kari Paul. *The rebel group stopping self-driving cars in their tracks – one cone at a time*. 2023. URL: <https://www.theguardian.com/us-news/2023/jul/26/san-francisco-stop-self-driving-cars-traffic-cone-safe-street-rebel>.
- [397] Billy Perrigo. *150 African Workers for ChatGPT, TikTok and Facebook Vote to Unionize at Landmark Nairobi Meeting*. 2023. URL: <https://time.com/6275995/chatgpt-facebook-african-workers-union/>.
- [398] Billy Perrigo. *Inside Facebook’s African Sweatshop*. 2022. URL: <https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/>.

- [399] James Pierce. “Undesigning technology: considering the negation of design by design”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '12. Austin, Texas, USA: Association for Computing Machinery, 2012, 957–966. ISBN: 9781450310154. DOI: [10.1145/2207676.2208540](https://doi.org/10.1145/2207676.2208540). URL: <https://doi.org/10.1145/2207676.2208540>.
- [400] Peter Pirolli and Stuart Card. “The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis”. In: *Proceedings of international conference on intelligence analysis*. Vol. 5. McLean, VA, USA. 2005, pp. 2–4.
- [401] Giada Pistilli. *HuggingFace announced the new feature to flag any Model, Dataset, or Space on the Hub*. 2022. URL: <https://twitter.com/GiadaPistilli/status/1571865167092396033?s=20&t=LRhhEu63s6ftPmtZdfz8Cw>.
- [402] Mikaela Pitcan, Alice E Marwick, and danah boyd. “Performing a Vanilla Self: Respectability Politics, Social Class, and the Digital World”. In: *Journal of Computer-Mediated Communication* 23.3 (Apr. 2018), pp. 163–179. ISSN: 1083-6101. DOI: [10.1093/jcmc/zmy008](https://doi.org/10.1093/jcmc/zmy008). eprint: <https://academic.oup.com/jcmc/article-pdf/23/3/163/24756585/zmy008.pdf>. URL: <https://doi.org/10.1093/jcmc/zmy008>.
- [403] Marie-Therese Png. “At the Tensions of South and North: Critical Roles of Global South Stakeholders in AI Governance”. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, 1434–1445. ISBN: 9781450393522. DOI: [10.1145/3531146.3533200](https://doi.org/10.1145/3531146.3533200). URL: <https://doi.org/10.1145/3531146.3533200>.
- [404] Jon Porter. *UK ditches exam results generated by biased algorithm after student protests*. 2020. URL: <https://www.theverge.com/2020/8/17/21372045/uk-a-level-results-algorithm-biased-coronavirus-covid-19-pandemic-university-applications>.
- [405] Eric A. Posner, E. Glen Weyl, Vitalik Buterin, and Jaron Lanier. “Data as Labor: VALUING INDIVIDUAL CONTRIBUTIONS TO THE DIGITAL ECONOMY”. In: *Radical Markets: Uprooting Capitalism and Democracy for a Just Society*. Princeton University Press, 2018, pp. 205–249. ISBN: 9780691196060. URL: <http://www.jstor.org/stable/j.ctvdf0kwg.11> (visited on 09/03/2024).
- [406] Neil Postman. *Five Things We Need to Know About Technological Change*. 1998. URL: <https://www.cs.ucdavis.edu/~rogaway/classes/188/materials/postman.pdf>.
- [407] Marcelo O. R. Prates, Pedro H. Avelar, and Luís C. Lamb. “Assessing gender bias in machine translation: a case study with Google Translate”. In: *Neural Computing and Applications* 32 (2019), 6363–6381. DOI: [10.1007/s00521-019-04144-6](https://doi.org/10.1007/s00521-019-04144-6). URL: <https://doi.org/10.1007/s00521-019-04144-6>.
- [408] Marcelo O. R. Prates, Pedro H. Avelar, and Luís C. Lamb. “Assessing gender bias in machine translation: a case study with Google Translate”. In: *Neural Computing and Applications* 32 (2020), 6363–6381. DOI: [10.1007/s00521-019-04144-6](https://doi.org/10.1007/s00521-019-04144-6). URL: <https://doi.org/10.1007/s00521-019-04144-6>.
- [409] Seeyam Quine. *Kristine Is Not Well: Memes vs. Censorship in metaverse inspired by real rebellion stories*. URL: <https://seeyam.com/kinw>.
- [410] Charles C. Ragin and Howard S. Becker. *What is a case?: Exploring the foundations of social inquiry*. Cambridge University Press, 1992.

- [411] Maxim Raginsky. @mraginsky. 2019. URL: <https://twitter.com/mraginsky/status/1080568937656565761>.
- [412] Maxim Raginsky. @mraginsky. 2019. URL: <https://twitter.com/mraginsky/status/1080569326644748289>.
- [413] Stephanie Carroll Rainie, Tahu Kukutai, Maggie Walter, Oscar Luis Figueroa-Rodríguez, Jennifer Walker, and Per Axelsson. "Indigenous data sovereignty". en. In: *The State of Open Data: Histories and Horizons*. 21. Cape Town and Ottawa: African Minds and the International Development Research Centre (IDRC), 2019, pp. 300–319. URL: <https://hdl.handle.net/10289/12918>.
- [414] 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*. AIES '19. Honolulu, HI, USA: Association for Computing Machinery, 2019, 429–435. ISBN: 9781450363242. DOI: [10.1145/3306618.3314244](https://doi.org/10.1145/3306618.3314244). URL: <https://doi.org/10.1145/3306618.3314244>.
- [415] Inioluwa Deborah Raji and Joy Buolamwini. "Actionable Auditing Revisited: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products". In: *Commun. ACM* 66.1 (Dec. 2022), 101–108. ISSN: 0001-0782. DOI: [10.1145/3571151](https://doi.org/10.1145/3571151). URL: <https://doi.org/10.1145/3571151>.
- [416] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. "Saving face: Investigating the ethical concerns of facial recognition auditing". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020, pp. 145–151.
- [417] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. "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*. FAT* '20. Barcelona, Spain: Association for Computing Machinery, 2020, 33–44. ISBN: 9781450369367. DOI: [10.1145/3351095.3372873](https://doi.org/10.1145/3351095.3372873). URL: <https://doi.org/10.1145/3351095.3372873>.
- [418] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. "Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance". In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 2022, pp. 557–571.
- [419] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. "Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices". In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1 (2021), pp. 1–23.
- [420] Divya Ramesh, Vaishnav Kameswaran, Ding Wang, and Nithya Sambasivan. "How Platform-User Power Relations Shape Algorithmic Accountability: A Case Study of Instant Loan Platforms and Financially Stressed Users in India". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, 1917–1928. ISBN: 9781450393522. DOI: [10.1145/3531146.3533237](https://doi.org/10.1145/3531146.3533237). URL: <https://doi.org/10.1145/3531146.3533237>.

- [421] Ryan Randazzo. *A slashed tire, a pointed gun, bullies on the road: Why do Waymo self-driving vans get so much hate?* 2018. URL: <https://www.azcentral.com/story/money/business/tech/2018/12/11/waymo-self-driving-vehicles-face-harassment-road-rage-phoenix-area/2198220002/>.
- [422] Noopur Raval, Rida Qadri, Richmond Y. Wong, Tamara Kneese, and Alex Hanna. "Considerations for Building Solidarity among Academic and Tech Workers: Thinking through Access, Positionality and Limits to Collective Action". In: *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI EA '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391566. DOI: [10.1145/3491101.3516511](https://doi.org/10.1145/3491101.3516511). URL: <https://doi.org/10.1145/3491101.3516511>.
- [423] *Regulation (EU) 2016/679 (General Data Protection Regulation)*. 2016.
- [424] Frederick Reiber, Nathan Chan-Yeong Kim, Allison McDonald, and Dana Calacci. "Surveillance, Spacing, Screaming and Scabbing: How Digital Technology Facilitates Union Busting". In: *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. CHI '26. New York, NY, USA: Association for Computing Machinery, 2026. ISBN: 9798400722783. DOI: [10.1145/3772318.3791507](https://doi.org/10.1145/3772318.3791507). URL: <https://doi.org/10.1145/3772318.3791507>.
- [425] Microsoft Research. *AI Fairness Checklist*. 2022. URL: <https://blog.google/technology/ai/join-us-in-the-ai-test-kitchen/>.
- [426] People + AI Research. "People AI Guidebook". In: (2021). URL: <https://pair.withgoogle.com/guidebook/>.
- [427] Ali Rezasab. @Arnasab. 2020. URL: <https://twitter.com/Arnasab/status/1290956206232698880>.
- [428] Andrea Rezzani, Antonella De Angeli, María Menéndez Blanco, and Max Dorfmann. "The space of user aggression in Human-Robot Interaction". In: *Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter*. CHI-taly '23. Torino, Italy: Association for Computing Machinery, 2023. ISBN: 9798400708060. DOI: [10.1145/3605390.3605402](https://doi.org/10.1145/3605390.3605402). URL: <https://doi.org/10.1145/3605390.3605402>.
- [429] Andrea Rezzani, María Menéndez-Blanco, Brad J. Bushman, and Antonella De Angeli. "What is robot abuse? A sociotechnical definition". In: *Behaviour & Information Technology* 0.0 (2025), pp. 1–19. DOI: [10.1080/0144929X.2025.2462268](https://doi.org/10.1080/0144929X.2025.2462268). eprint: <https://doi.org/10.1080/0144929X.2025.2462268>. URL: <https://doi.org/10.1080/0144929X.2025.2462268>.
- [430] Brianna Richardson, Jean Garcia-Gathright, Samuel F. Way, Jennifer Thom, and Henriette Cramer. "Towards Fairness in Practice: A Practitioner-Oriented Rubric for Evaluating Fair ML Toolkits". 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: [10.1145/3411764.3445604](https://doi.org/10.1145/3411764.3445604). URL: <https://doi.org/10.1145/3411764.3445604>.
- [431] Ronald E. Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. "Auditing Partisan Audience Bias within Google Search". In: *Proc. ACM Hum.-Comput. Interact.* 2.CSCW (Nov. 2018). DOI: [10.1145/3274417](https://doi.org/10.1145/3274417). URL: <https://doi.org/10.1145/3274417>.

- [432] Samantha Robertson and Niloufar Salehi. "What If I Don't Like Any Of The Choices? The Limits of Preference Elicitation for Participatory Algorithm Design". In: *Workshop on Participatory Approaches to Machine Learning at ICML 2020*. 2020.
- [433] Aja Romano. "A group of YouTubers is trying to prove the site systematically demonetizes queer content". In: *Vox* (Oct. 2019). URL: <https://www.vox.com/culture/2019/10/10/20893258/youtube-lgbtq-censorship-demonetization-nerd-city-algorithm-report>.
- [434] Simon Romero. *Wielding Rocks and Knives, Arizonans Attack Self-Driving Cars*. 2018. URL: <https://www.nytimes.com/2018/12/31/us/waymo-self-driving-cars-arizona-attacks.html>.
- [435] Jon Roozenbeek, Melisa Basol, and Sander van der Linden. *A New Way to Inoculate People Against Misinformation*. 2021. URL: <https://behavioralscientist.org/a-new-way-to-inoculate-people-against-misinformation/>.
- [436] Jean-Jacques Rousseau. *The Social Contract and Discourse on the Origin of Inequality*. Ed. by Lester G. Crocker. Washington Square Press, 1967.
- [437] Anita Roychoudhury, Daniel P. Shepardson, Andrew Hirsch, Devdutta Niyogi, Jignesh Mehta, and Sara Top. "The Need to Introduce System Thinking in Teaching Climate Change". In: *Science Educator*. 2nd ed. Vol. 25. National Science Education Leadership Association, 2017, 73–81.
- [438] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. "Imagenet large scale visual recognition challenge". In: *International journal of computer vision* 115.3 (2015), pp. 211–252.
- [439] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. "The cost structure of sensemaking". In: (1993), pp. 269–276.
- [440] Dan S. *Yelp Routinely "Filters" My Customer Reviews Because I won't advertise with them*. 2013. URL: <https://www.yelp.com/topic/denver-yelp-routinely-filters-my-customer-reviews-because-i-wont-advertise-with-them>.
- [441] Samar Sabie, Robert Soden, Steven Jackson, and Tapan Parikh. "Unmaking as Emancipation: Lessons and Reflections from Luddism". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3581412. URL: <https://doi.org/10.1145/3544548.3581412>.
- [442] Henrik Skaug Sætra. "A shallow defence of a technocracy of artificial intelligence: Examining the political harms of algorithmic governance in the domain of government". In: *Technology in Society* 62 (2020), p. 101283. ISSN: 0160-791X. DOI: <https://doi.org/10.1016/j.techsoc.2020.101283>. URL: <https://www.sciencedirect.com/science/article/pii/S0160791X19305925>.
- [443] Niloufar Salehi, Lilly C. Irani, Michael S. Bernstein, Ali Alkhatib, Eva Ogbé, Kristy Milland, and Clickhappier. "We Are Dynamo: Overcoming Stalling and Friction in Collective Action for Crowd Workers". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. Seoul, Republic of Korea: Association for Computing Machinery, 2015, 1621–1630. ISBN: 9781450331456. DOI: 10.1145/2702123.2702508. URL: <https://doi.org/10.1145/2702123.2702508>.

- [444] Niloufar Salehi, Roya Pakzad, Nazita Lajevardi, and Mariam Asad. "Sustained Harm Over Time and Space Limits the External Function of Online Counterpublics for American Muslims". In: 6.CSCW1 (2023). DOI: [10.1145/3579526](https://doi.org/10.1145/3579526). URL: <https://doi.org/10.1145/3579526>.
- [445] Eryk Salvaggio. *Slop Infrastructures 1 & 2*. 2024. URL: <https://mail.cyberneticforests.com/slop-infrastructures-1-2/>.
- [446] P. Salvini, G. Ciaravella, W. Yu, G. Ferri, A. Manzi, B. Mazzolai, C. Laschi, S.R. Oh, and P. Dario. "How safe are service robots in urban environments? Bullying a robot". In: *19th International Symposium in Robot and Human Interactive Communication*. 2010, pp. 1–7. DOI: [10.1109/ROMAN.2010.5654677](https://doi.org/10.1109/ROMAN.2010.5654677).
- [447] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. "Auditing algorithms: Research methods for detecting discrimination on internet platforms". In: *Data and Discrimination: Converting Critical Concerns into Productive Inquiry* (2014). URL: <https://social.cs.uiuc.edu/papers/pdfs/ICA2014-Sandvig.pdf>.
- [448] Satviki Sanjay. "Twitter's AI Revealed to Have Racial Bias After a Viral Photo Experiment". In: *Vice* (2020). URL: <https://www.vice.com/en/article/93547v/twitter-ai-algorithm-has-racial-bias>.
- [449] Aaron Sankin, Dhruv Mehrotra, Surya Mattu, and Annie Gilbertson. *Crime Prediction Software Promised to Be Free of Biases. New Data Shows It Perpetuates Them*. 2021. URL: <https://themarkup.org/prediction-bias/2021/12/02/crime-prediction-software-promised-to-be-free-of-biases-new-data-shows-it-perpetuates-them>.
- [450] Chris Sawyer. *RollerCoaster Tycoon*. 1999. URL: <https://atari.com/pages/rollercoaster-tycoon>.
- [451] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. "Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work". In: *Proc. ACM Hum.-Comput. Interact.* 2.CSCW (2018). DOI: [10.1145/3274423](https://doi.org/10.1145/3274423). URL: <https://doi.org/10.1145/3274423>.
- [452] Stuart Schechter. "How to buy better testing using competition to get the most security and robustness for your dollar". In: *International Conference on Infrastructure Security*. Springer. 2002, pp. 73–87.
- [453] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R. Brubaker. "A Framework of Severity for Harmful Content Online". In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW2 (2021). DOI: [10.1145/3479512](https://doi.org/10.1145/3479512). URL: <https://doi.org/10.1145/3479512>.
- [454] Morgan Klaus Scheuerman, Jacob M. Paul, and Jed R. Brubaker. "How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services". In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). DOI: [10.1145/3359246](https://doi.org/10.1145/3359246). URL: <https://doi.org/10.1145/3359246>.
- [455] Morgan Klaus Scheuerman, Katta Spiel, Oliver L. Haimson, Foad Hamidi, and Stacy M. Branham. *HCI Guidelines for Gender Equity and Inclusivity*. 2019. URL: <https://www.morgan-klaus.com/gender-guidelines.html>.

- [456] Hanna Schneider, Malin Eiband, Daniel Ullrich, and Andreas Butz. "Empowerment in HCI - A Survey and Framework". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal QC, Canada: Association for Computing Machinery, 2018, 1–14. ISBN: 9781450356206. DOI: 10.1145/3173574.3173818. URL: <https://doi.org/10.1145/3173574.3173818>.
- [457] Sarita Schoenebeck. "Example demographic questions for surveys or experiments". In: (). URL: <https://twitter.com/syardi/status/1276587095767887874>.
- [458] Oscar Schwartz. *Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation The bot learned language from people on Twitter—but it also learned values*. 2019. URL: <https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation>.
- [459] Carol F Scott, Gabriela Marcu, Riana Elyse Anderson, Mark W Newman, and Sarita Schoenebeck. "Trauma-Informed Social Media: Towards Solutions for Reducing and Healing Online Harm". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3581512. URL: <https://doi.org/10.1145/3544548.3581512>.
- [460] James C. Scott. *Weapons of the Weak: Everyday Forms of Peasant Resistance*. Yale University Press, 1985, pp. xv–xxii. ISBN: 9780300036411. URL: <http://www.jstor.org/stable/j.ctt1nq836>.
- [461] "Search Engine Bias". In: (2019). URL: <https://therepproject.org/search-engine-bias/>.
- [462] Nick Seaver. "Algorithms as culture: Some tactics for the ethnography of algorithmic systems". In: *Big Data & Society* 4.2 (2017), p. 2053951717738104. DOI: 10.1177/2053951717738104. eprint: <https://doi.org/10.1177/2053951717738104>. URL: <https://doi.org/10.1177/2053951717738104>.
- [463] Nick Seaver. "Knowing algorithms". In: *Digital STS* (2019), pp. 412–422.
- [464] Joseph Seering, Felicia Ng, Zheng Yao, and Geoff Kaufman. "Applications of Social Identity Theory to Research and Design in Computer-Supported Cooperative Work". In: *Proc. ACM Hum.-Comput. Interact.* 2.CSCW (2018). DOI: 10.1145/3274771. URL: <https://doi.org/10.1145/3274771>.
- [465] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. "Fairness and Abstraction in Sociotechnical Systems". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Atlanta, GA, USA: Association for Computing Machinery, 2019, 59–68. ISBN: 9781450361255. DOI: 10.1145/3287560.3287598. URL: <https://doi.org/10.1145/3287560.3287598>.
- [466] Bryan Semaan. "'Routine Infrastructuring' as 'Building Everyday Resilience with Technology': When Disruption Becomes Ordinary". In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (2019). DOI: 10.1145/3359175. URL: <https://doi.org/10.1145/3359175>.
- [467] Farhana Shahid and Aditya Vashistha. "Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony?" In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3581538. URL: <https://doi.org/10.1145/3544548.3581538>.

- [468] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. "Glaze: protecting artists from style mimicry by text-to-image models". In: *Proceedings of the 32nd USENIX Conference on Security Symposium*. SEC '23. Anaheim, CA, USA: USENIX Association, 2023. ISBN: 978-1-939133-37-3.
- [469] Shawn Shan, Wenxin Ding, Josephine Passananti, Stanley Wu, Haitao Zheng, and Ben Y. Zhao. "Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models". In: *2024 IEEE Symposium on Security and Privacy (SP)*. 2024, pp. 807–825. DOI: [10.1109/SP54263.2024.00207](https://doi.org/10.1109/SP54263.2024.00207).
- [470] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y. Zhao. "Fawkes: Protecting Privacy against Unauthorized Deep Learning Models". In: *29th USENIX Security Symposium (USENIX Security 20)*. USENIX Association, Aug. 2020, pp. 1589–1604. ISBN: 978-1-939133-17-5. URL: <https://www.usenix.org/conference/usenixsecurity20/presentation/shan>.
- [471] Tanusree Sharma, Smirity Kaushik, Yaman Yu, Syed Ishtiaque Ahmed, and Yang Wang. "User Perceptions and Experiences of Targeted Ads on Social Media Platforms: Learning from Bangladesh and India". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: [10.1145/3544548.3581498](https://doi.org/10.1145/3544548.3581498). URL: <https://doi.org/10.1145/3544548.3581498>.
- [472] Sam Shead. *How a computer algorithm caused a grading crisis in British schools*. 2020. URL: <https://www.cnn.com/2020/08/21/computer-algorithm-caused-a-grading-crisis-in-british-schools.html>.
- [473] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Roshtamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. "Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction". In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '23. Montréal, QC, Canada: Association for Computing Machinery, 2023, 723–741. ISBN: 9798400702310. DOI: [10.1145/3600211.3604673](https://doi.org/10.1145/3600211.3604673). URL: <https://doi.org/10.1145/3600211.3604673>.
- [474] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. "Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors". In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW2 (2021). DOI: [10.1145/3479577](https://doi.org/10.1145/3479577). URL: <https://doi.org/10.1145/3479577>.
- [475] Hong Shen, Leijie Wang, Wesley H Deng, Ciell Brusse, Ronald Velgersdijk, and Haiyi Zhu. "The Model Card Authoring Toolkit: Toward Community-centered, Deliberation-driven AI Design". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022, pp. 440–451.
- [476] Limor Shifman. "Memes in a digital world: Reconciling with a conceptual troublemaker". In: *Journal of Computer-Mediated Communication* 18.3 (2013), pp. 362–377.
- [477] Andrew Shouldice. *Tunic*. 2022. URL: <https://tunicgame.com>.

- [478] Valerie J. Shute and Fengfeng Ke. "Games, Learning, and Assessment". In: *Assessment in Game-Based Learning: Foundations, Innovations, and Perspectives*. Ed. by Dirk Ifenthaler, Deniz Eseryel, and Xun Ge. New York, NY: Springer New York, 2012, pp. 43–58. ISBN: 978-1-4614-3546-4. DOI: [10.1007/978-1-4614-3546-4_4](https://doi.org/10.1007/978-1-4614-3546-4_4). URL: https://doi.org/10.1007/978-1-4614-3546-4_4.
- [479] SIGCHI's Involvement in Public Policy. 2023. URL: <https://sigchi.org/public-policy/>.
- [480] Ignacio Siles, Andrés Segura-Castillo, Ricardo Solís, and Mónica Sancho. "Folk theories of algorithmic recommendations on Spotify: Enacting data assemblages in the global South". In: *Big Data & Society* 7.1 (2020), p. 2053951720923377. DOI: [10.1177/2053951720923377](https://doi.org/10.1177/2053951720923377). eprint: <https://doi.org/10.1177/2053951720923377>. URL: <https://doi.org/10.1177/2053951720923377>.
- [481] Kalina Silverman. *Big Talk*. 2014. URL: <https://makebigtalk.app>.
- [482] Jordan Simonovski. @_jsimonovski: "I wonder if Twitter does this to fictional characters too." with related images labelled "Lenny" and "Carl". 2020. URL: https://twitter.com/_jsimonovski/status/1307542747197239296.
- [483] Audra Simpson. "On Ethnographic Refusal: Indigeneity, 'Voice' and Colonial Citizenship". In: *Junctures: The Journal for Thematic Dialogue* (2007), pp. 67–80.
- [484] Audra Simpson. "The ruse of consent and the anatomy of 'refusal': cases from indigenous North America and Australia". In: *Postcolonial Studies* 20.1 (2017), pp. 18–33. DOI: [10.1080/13688790.2017.1334283](https://doi.org/10.1080/13688790.2017.1334283). eprint: <https://doi.org/10.1080/13688790.2017.1334283>. URL: <https://doi.org/10.1080/13688790.2017.1334283>.
- [485] Ellen Simpson and Bryan Semaan. "For You, or For 'You'? Everyday LGBTQ+ Encounters with TikTok". In: *Proc. ACM Hum.-Comput. Interact.* 4.CSCW3 (Jan. 2021). DOI: [10.1145/3432951](https://doi.org/10.1145/3432951). URL: <https://doi.org/10.1145/3432951>.
- [486] Anubha Singh and Tina Park. "Automating Care: Online Food Delivery Work During the CoVID-19 Crisis in India". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, 160–172. ISBN: 9781450393522. DOI: [10.1145/3531146.3533082](https://doi.org/10.1145/3531146.3533082). URL: <https://doi.org/10.1145/3531146.3533082>.
- [487] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. "Participation Is not a Design Fix for Machine Learning". In: *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. EAAMO '22. Arlington, VA, USA: Association for Computing Machinery, 2022. ISBN: 9781450394772. DOI: [10.1145/3551624.3555285](https://doi.org/10.1145/3551624.3555285). URL: <https://doi.org/10.1145/3551624.3555285>.
- [488] Media Smarts. #ForYou: A Game About Algorithms. URL: <https://mediasmarts.ca/digital-media-literacy/educational-games/foryou-game-about-algorithms>.
- [489] Jack Smith IV. (Exclusive) Crime-prediction tool PredPol amplifies racially biased policing, study shows. 2016. URL: <https://www.mic.com/articles/156286/crime-prediction-tool-pred-pol-only-amplifies-racially-biased-policing-study-shows>.

- [490] Rion Snow, Brendan O'Connor, Dan Jurafsky, and Andrew Y Ng. "Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks". In: *Proceedings of the 2008 conference on empirical methods in natural language processing*. 2008, pp. 254–263.
- [491] Katherine W Song and Eric Paulos. "Unmaking: Enabling and Celebrating the Creative Material of Failure, Destruction, Decay, and Deformation". 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: [10.1145/3411764.3445529](https://doi.org/10.1145/3411764.3445529). URL: <https://doi.org/10.1145/3411764.3445529>.
- [492] Katherine W. Song, Samar Sabie, Steven J. Jackson, Kristina Lindström, Eric Paulos, Åsa Ståhl, and Ron Wakkary. "Unmaking and HCI: Techniques, Technologies, Materials, and Philosophies beyond Making". In: *ACM Trans. Comput.-Hum. Interact.* 31.6 (Jan. 2025). ISSN: 1073-0516. DOI: [10.1145/3689047](https://doi.org/10.1145/3689047). URL: <https://doi.org/10.1145/3689047>.
- [493] Anne Spaa, Abigail Durrant, Chris Elsdén, and John Vines. "Understanding the Boundaries between Policymaking and HCI". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, 1–15. ISBN: 9781450359702. DOI: [10.1145/3290605.3300314](https://doi.org/10.1145/3290605.3300314). URL: <https://doi.org/10.1145/3290605.3300314>.
- [494] Spenser Starke. *Alice is Missing*. 2020. URL: <https://www.huntersentertainment.com/alice-is-missing>.
- [495] StarshineScribbles. *'Twas The Labour Dispute Before Christmas*. URL: <https://starshinescribbles.itch.io/labourdispute>.
- [496] *Statement on Supreme Court Decision Removing Hurdles to Online Civil Rights Testing and Research*. 2021. URL: <https://www.aclu.org/press-releases/statement-supreme-court-decision-removing-hurdles-online-civil-rights-testing-and>.
- [497] Jaime Stathis. *What Is Mastodon, and Why Are People Leaving Twitter for It?* 2023. URL: <https://www.rd.com/article/what-is-mastodon/>.
- [498] Ella Steen, Kathryn Yurechko, and Daniel Klug. "You Can (Not) Say What You Want: Using Algospeak to Contest and Evade Algorithmic Content Moderation on TikTok". In: *Social Media + Society* 9.3 (2023), p. 20563051231194586. DOI: [10.1177/20563051231194586](https://doi.org/10.1177/20563051231194586). eprint: <https://doi.org/10.1177/20563051231194586>. URL: <https://doi.org/10.1177/20563051231194586>.
- [499] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. "The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support". In: *Proceedings of the 2021 CHI conference on human factors in computing systems*. 2021, pp. 1–14.
- [500] Norman Makoto Su, Amanda Lazar, and Lilly Irani. "Critical Affects: Tech Work Emotions Amidst the Techlash". In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW1 (Apr. 2021). DOI: [10.1145/3449253](https://doi.org/10.1145/3449253). URL: <https://doi.org/10.1145/3449253>.
- [501] Peter Suber. *Nomic*. 1982. URL: <https://boardgamegeek.com/boardgame/14451/nomic>.

- [502] Lucy Suchman. "Located accountabilities in technology production". In: *Scand. J. Inf. Syst.* 14.2 (Sept. 2002), 91–105. ISSN: 0905-0167.
- [503] Lucy Suchman. "The uncontroversial 'thingness' of AI". In: *Big Data & Society* 10.2 (2023), p. 20539517231206794. DOI: [10.1177/20539517231206794](https://doi.org/10.1177/20539517231206794). eprint: <https://doi.org/10.1177/20539517231206794>. URL: <https://doi.org/10.1177/20539517231206794>.
- [504] Lucy A Suchman. *Plans and situated actions: The problem of human-machine communication*. Cambridge University Press, 1987.
- [505] Jina Suh, Soroush Ghorashi, Gonzalo Ramos, Nan-Chen Chen, Steven Drucker, Johan Verwey, and Patrice Simard. "AnchorViz: Facilitating Semantic Data Exploration and Concept Discovery for Interactive Machine Learning". In: *ACM Trans. Interact. Intell. Syst.* 10.1 (Aug. 2019). ISSN: 2160-6455. DOI: [10.1145/3241379](https://doi.org/10.1145/3241379). URL: <https://doi.org/10.1145/3241379>.
- [506] Cella M. Sum and Alicia DeVrio. "Shifting Power Through Resistance: Social Justice in Practice". In: *XRDS* 30.4 (June 2024), 7–9. ISSN: 1528-4972. DOI: [10.1145/3665272](https://doi.org/10.1145/3665272). URL: <https://doi.org/10.1145/3665272>.
- [507] Cella M. Sum, Anna Konvicka, Mona Wang, and Sarah E. Fox. "The Future of Tech Labor: How Workers are Organizing and Transforming the Computing Industry". In: *Proc. ACM Hum.-Comput. Interact.* 9.7 (Oct. 2025). DOI: [10.1145/3757671](https://doi.org/10.1145/3757671). URL: <https://doi.org/10.1145/3757671>.
- [508] Cella M Sum, Anh-Ton Tran, Jessica Lin, Rachel Kuo, Cynthia L Bennett, Christina Harrington, and Sarah E Fox. "Translation as (Re)Mediation: How Ethnic Community-Based Organizations Negotiate Legitimacy". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: [10.1145/3544548.3581280](https://doi.org/10.1145/3544548.3581280). URL: <https://doi.org/10.1145/3544548.3581280>.
- [509] Ping Sun. "Your order, their labor: An exploration of algorithms and laboring on food delivery platforms in China". In: *Chinese Journal of Communication* 12.3 (2019), pp. 308–323. DOI: [10.1080/17544750.2019.1583676](https://doi.org/10.1080/17544750.2019.1583676). eprint: <https://doi.org/10.1080/17544750.2019.1583676>. URL: <https://doi.org/10.1080/17544750.2019.1583676>.
- [510] Harini Suresh, Rajiv Movva, Amelia Lee Dogan, Rahul Bhargava, Isadora Cruxen, Angeles Martinez Cuba, Guilia Taurino, Wonyoung So, and Catherine D'Ignazio. "Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Femicide Counterdata Collection". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, 667–678. ISBN: 9781450393522. DOI: [10.1145/3531146.3533132](https://doi.org/10.1145/3531146.3533132). URL: <https://doi.org/10.1145/3531146.3533132>.
- [511] Office of the Surgeon General. *Social Media and Youth Mental Health*. 2023. URL: <https://www.hhs.gov/surgeongeneral/priorities/youth-mental-health/social-media/index.html>.
- [512] Latanya Sweeney. "Discrimination in Online Ad Delivery: Google ads, black names and white names, racial discrimination, and click advertising". In: *Queue* 11.3 (Mar. 2013), 10–29. ISSN: 1542-7730. DOI: [10.1145/2460276.2460278](https://doi.org/10.1145/2460276.2460278). URL: <https://doi.org/10.1145/2460276.2460278>.

- [513] Sam Sweeney. *Uber, Lyft drivers manipulate fares at Reagan National causing artificial price surges*. 2019. URL: <https://wjla.com/news/local/uber-and-lyft-drivers-fares-at-reagan-national>.
- [514] Jordan Taylor, Joel Mire, Alicia DeVrio, Maarten Sap, Haiyi Zhu, and Sarah E. Fox. "I Just Don't Want My Work Being Fed Into The AI Blender": Queer Artists on Refusing and Resisting Generative AI. 2026. arXiv: 2604.14266 [cs.HC]. URL: <https://arxiv.org/abs/2604.14266>.
- [515] The Waymo Team. *Independent Audits of Waymo's Safety Case and Remote Assistance Programs*. 2025. URL: <https://waymo.com/blog/2025/11/independent-audits/>.
- [516] Arvi Teikari. *Baba Is You*. 2019. URL: <https://www.hempuli.com/baba/>.
- [517] Alexandra Teixeira Riggs and Noura Howell. "Reconfiguring through Ruptures: Material Reconfigurations and Un/Making as Tangible Tactics for Queering AI-Generated Histories". In: *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. CHI '26. New York, NY, USA: Association for Computing Machinery, 2026. ISBN: 9798400722783. DOI: 10.1145/3772318.3791302. URL: <https://doi.org/10.1145/3772318.3791302>.
- [518] Tiziana Terranova. "Free Labor: PRODUCING CULTURE FOR THE DIGITAL ECONOMY". In: *Social Text* 18.2 (63) (June 2000), pp. 33–58. ISSN: 0164-2472. DOI: 10.1215/01642472-18-2_63-33. eprint: [https://read.dukeupress.edu/social-text/article-pdf/18/2\(63\)/33/513021/st_18_2-2.terranova.pdf](https://read.dukeupress.edu/social-text/article-pdf/18/2(63)/33/513021/st_18_2-2.terranova.pdf). URL: https://doi.org/10.1215/01642472-18-2_63-33.
- [519] Petros Terzis, Michael Veale, and Noëlle Gaumann. "Law and the Emerging Political Economy of Algorithmic Audits". In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '24. Rio de Janeiro, Brazil: Association for Computing Machinery, 2024, 1255–1267. ISBN: 9798400704505. DOI: 10.1145/3630106.3658970. URL: <https://doi.org/10.1145/3630106.3658970>.
- [520] Klaus Teuber. *Catan*. 1995. URL: <https://www.catan.com>.
- [521] Lucas Theis and Zehan Wang. *Speedy Neural Networks for Smart Auto-Cropping of Images*. 2018. URL: https://blog.twitter.com/engineering/en_us/topics/infrastructure/2018/Smart-Auto-Cropping-of-Images.html.
- [522] Julien Thiennot. *Cookie CLicker*. 2013. URL: <https://cookieclicker.com>.
- [523] Alexandra To, Hillary Carey, Riya Shrivastava, Jessica Hammer, and Geoff Kaufman. "Interactive Fiction Provotypes for Coping with Interpersonal Racism". 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: 10.1145/3491102.3502044. URL: <https://doi.org/10.1145/3491102.3502044>.
- [524] Eve Tuck and K. Wayne Yang. "R-Words: Refusing Research". In: *Humanizing research: Decolonizing qualitative inquiry with youth and communities* 223 (Jan. 2014), pp. 223–248. DOI: 10.4135/9781544329611.n12. URL: <https://static1.squarespace.com/static/557744ffe4b013bae3b7af63/t/557f2ee5e4b0220eff4ae4b5/1434398437409/Tuck+and+Yang+R+Words+Refusing+Research.pdf>.

- [525] Harry Tucker. "Australian Uber drivers say the company is manipulating their ratings to boost its fees". In: *Businessinsider* (May 2016). URL: <https://www.businessinsider.com.au/australian-uber-drivers-say-the-company-is-manipulating-their-ratings-to-boost-the-companys-fees-2016-5>.
- [526] Moyra Turkington. *Against the Grain*. URL: <https://unrulydesigns.com/war-birds/against-the-grain-by-moyra-turkington>.
- [527] Amos Tversky and Daniel Kahneman. "Loss Aversion in Riskless Choice: A Reference-Dependent Model". In: *The Quarterly Journal of Economics* 106.4 (1991), pp. 1039–1061. ISSN: 00335533, 15314650. URL: <http://www.jstor.org/stable/2937956> (visited on 09/12/2024).
- [528] "Twitter investigates racial bias in image previews". In: *BBC News* (2020). URL: <https://www.bbc.com/news/technology-54234822>.
- [529] *Twitter thread*. Last Friday, CBS cancelled a segment about our clients suing Amazon for selling suicide kits to their now deceased kids. CBS' cowardice gave me renewed clarity about how urgent this litigation is. 1/. 2022. URL: <https://twitter.com/cagoldberglaw/status/1578121292502409216>.
- [530] Megan Twohey and Gabriel J.X. Dance. *Lawmakers Press Amazon on Sales of Chemical Used in Suicides*. 2022. URL: <https://www.nytimes.com/2022/02/04/technology/amazon-suicide-poison-preservative.html>.
- [531] Alec Tyson and Emma Kikuchi. *Growing public concern about the role of artificial intelligence in daily life*. 2023. URL: <https://www.pewresearch.org/short-reads/2023/08/28/growing-public-concern-about-the-role-of-artificial-intelligence-in-daily-life/>.
- [532] Sohini Upadhyay, Dasha Pruss, Alicia DeVrio, Krzysztof Z. Gajos, and Naveena Karusala. "'It just requires so much more creativity': Barriers and Workarounds to Gathering Information for AI Contestation". In: *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. CHI '26. New York, NY, USA: Association for Computing Machinery, 2026. ISBN: 9798400722783. DOI: 10.1145/3772318.3791923. URL: <https://doi.org/10.1145/3772318.3791923>.
- [533] Steffie de Vaan. *Bitches in the Club*. URL: <https://steffie-de-vaan.itch.io/bitches-in-the-club>.
- [534] Dominique A. Montiel Valle and Samantha Shorey. "Dis//assemblages of AI: repair labor and resistance in the automated workplace". In: *Information, Communication & Society* 27.10 (2024), pp. 2022–2037. DOI: 10.1080/1369118X.2024.2371794. eprint: <https://doi.org/10.1080/1369118X.2024.2371794>. URL: <https://doi.org/10.1080/1369118X.2024.2371794>.
- [535] Colin Vandenhof and Edith Law. "Contradict the Machine: A Hybrid Approach to Identifying Unknown Unknowns". In: *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 2019, pp. 2238–2240.
- [536] Dora Vargha. @DoraVargha: "Hungarian is a gender neutral language, it has no gendered pronouns, so Google Translate automatically chooses the gender for you. Here is how everyday sexism is consistently encoded in 2021. Fuck you, Google." with a related image. 2021. URL: <https://twitter.com/DoraVargha/status/1373211762108076034>.

- [537] Kush R Varshney. "Trustworthy machine learning and artificial intelligence". In: *XRDS: Crossroads, The ACM Magazine for Students* 25.3 (2019), pp. 26–29.
- [538] Jennifer Wortman Vaughan. "Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research." In: *J. Mach. Learn. Res.* 18.1 (2017), pp. 7026–7071.
- [539] Michael Veale and Reuben Binns. "Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data". In: *Big Data & Society* 4.2 (2017), p. 2053951717743530.
- [540] Michael Veale, Max Van Kleek, and Reuben Binns. "Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal QC, Canada: Association for Computing Machinery, 2018, 1–14. ISBN: 9781450356206. DOI: [10.1145/3173574.3174014](https://doi.org/10.1145/3173574.3174014). URL: <https://doi.org/10.1145/3173574.3174014>.
- [541] Briana Vecchione, Karen Levy, and Solon Barocas. "Algorithmic Auditing and Social Justice: Lessons from the History of Audit Studies". In: *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. EAAMO '21. -, NY, USA: Association for Computing Machinery, 2021. ISBN: 9781450385534. DOI: [10.1145/3465416.3483294](https://doi.org/10.1145/3465416.3483294). URL: <https://doi.org/10.1145/3465416.3483294>.
- [542] Julia Velkova and Anne Kaun. "Algorithmic resistance: media practices and the politics of repair". In: *Information, Communication & Society* 24.4 (2021), pp. 523–540. DOI: [10.1080/1369118X.2019.1657162](https://doi.org/10.1080/1369118X.2019.1657162). eprint: <https://doi.org/10.1080/1369118X.2019.1657162>. URL: <https://doi.org/10.1080/1369118X.2019.1657162>.
- [543] Quentin Velluet. *Can Africa salvage its digital sovereignty?* 2021. URL: <https://www.theafricareport.com/80606/can-africa-salvage-its-digital-sovereignty/>.
- [544] Janet Vertesi. *About the Opt Out Project*. 2022. URL: <https://www.optoutproject.net/about-the-opt-out-project/>.
- [545] Janet Vertesi, danah boyd, Alex Taylor, and Benjamin Shestakofsky. *Reckoning with the Political Economy of AI: Avoiding Decoys in Pursuit of Accountability*. 2026. arXiv: [2604.16106](https://arxiv.org/abs/2604.16106) [cs.CY]. URL: <https://arxiv.org/abs/2604.16106>.
- [546] Neil Vigdor. "Apple card investigated after gender discrimination complaints". In: *The New York Times* (Nov. 2019). URL: <https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html>.
- [547] Salomé Viljoen. *Data as property?* 2020. URL: <https://www.phenomenalworld.org/analysis/data-as-property/>.
- [548] James Vincent. "Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech". In: *The Verge* (Jan. 2018). URL: <https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai>.
- [549] James Vincent. "Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day". In: *The Verge* (2016). URL: <https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist>.

- [550] Nicholas Vincent and Brent Hecht. "Can "Conscious Data Contribution" Help Users to Exert "Data Leverage" Against Technology Companies?" In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW1 (2021). DOI: [10.1145/3449177](https://doi.org/10.1145/3449177). URL: <https://doi.org/10.1145/3449177>.
- [551] Nicholas Vincent, Brent Hecht, and Shilad Sen. ""Data Strikes": Evaluating the Effectiveness of a New Form of Collective Action Against Technology Companies". In: *The World Wide Web Conference*. WWW '19. San Francisco, CA, USA: Association for Computing Machinery, 2019, 1931–1943. ISBN: 9781450366748. DOI: [10.1145/3308558.3313742](https://doi.org/10.1145/3308558.3313742). URL: <https://doi.org/10.1145/3308558.3313742>.
- [552] Nicholas Vincent, Hanlin Li, Nicole Tilly, Stevie Chancellor, and Brent Hecht. "Data Leverage: A Framework for Empowering the Public in Its Relationship with Technology Companies". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 2021, 215–227. ISBN: 9781450383097. DOI: [10.1145/3442188.3445885](https://doi.org/10.1145/3442188.3445885). URL: <https://doi.org/10.1145/3442188.3445885>.
- [553] Swen Vincke, David Walgrave, Bert Van Semmertier, Blake Maltby, Joachim Vleminckx, and Adam Smith. *Baldur's Gate 3*. 2023. URL: <https://baldursgate3.game>.
- [554] Lee Vinsel. *You're Doing It Wrong: Notes on Criticism and Technology Hype*. 2021. URL: <https://sts-news.medium.com/youre-doing-it-wrong-notes-on-criticism-and-technology-hype-18b08b4307e5>.
- [555] Daisuke Wakabayashi. *California Passes Sweeping Law to Protect Online Privacy*. 2018. URL: <https://www.nytimes.com/2018/06/28/technology/california-online-privacy-law.html>.
- [556] Daisuke Wakabayashi. *Silicon Valley Faces Regulatory Fight on Its Home Turf*. 2018. URL: <https://www.nytimes.com/2018/05/13/business/california-data-privacy-ballot-measure.html>.
- [557] Shuaishuai Wang. "Calculating dating goals: data gaming and algorithmic sociality on Blued, a Chinese gay dating app". In: *Information, Communication & Society* 23.2 (2020), pp. 181–197.
- [558] Xiaowei R. Wang, Qianqian Ye, and Jasmine Wang. *The Future of Memory, 2019-2022*. 2019-2022. URL: <https://www.xiaoweiwang.com/projects/new-portfolio-item>.
- [559] Zehan Wang. @ZehanWang: "We'll look into this. The algorithm does not do face detection at all (it actually replaced a previous algorithm which did). We conducted some bias studies before release back in 2017. At the time we found that there was no significant bias between ethnicities (or genders)." in response to quoted tweet: "Based on some experiments I tried, I think @colinmadland's facial hair is affecting the model because of the contrast with his skin. I removed his facial hair and the Black man shows in the preview for me. Our team did test for racial bias before shipping the model." 2020. URL: <https://twitter.com/ZehanWang/status/1307461285811032066>.
- [560] Tris Warkentin and Josh Woodward. *AI Test Kitchen*. 2022. URL: <https://blog.google/technology/ai/join-us-in-the-ai-test-kitchen/>.

- [561] Charlie Warzel. *The NAACP Wants Users To Log Out Of Facebook In Protest*. 2018. URL: <https://www.buzzfeednews.com/article/charliewarzel/naacp-log-out-facebook-protest>.
- [562] *We can't let Facebook do this again*. 2019. URL: <https://naacpculpeper.org/we-cant-let-facebook-do-this-again/>.
- [563] Cole Wehrle and Kyle Ferrin. *Root*. 2018. URL: <https://ledergames.com/products/root-a-game-of-woodland-might-and-right>.
- [564] Karl E Weick. *Sensemaking in organizations*. Vol. 3. Sage, 1995.
- [565] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. "Taxonomy of Risks Posed by Language Models". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, 214–229. ISBN: 9781450393522. DOI: [10.1145/3531146.3533088](https://doi.org/10.1145/3531146.3533088). URL: <https://doi.org/10.1145/3531146.3533088>.
- [566] Garrison Wells, Agnes Romhanyi, Jason G. Reitman, Reginald Gardner, Kurt Squire, and Constance Steinkuehler. "Right-Wing Extremism in Mainstream Games: A Review of the Literature". In: *Games and Culture* 19.4 (2024), pp. 469–492. DOI: [10.1177/15554120231167214](https://doi.org/10.1177/15554120231167214). eprint: <https://doi.org/10.1177/15554120231167214>. URL: <https://doi.org/10.1177/15554120231167214>.
- [567] Alex White. *Review - The Price of Coal RPG*. 2024. URL: <https://planesailinggames.com/review/2024-01-25-review-price-of-coal>.
- [568] Cedric Deslandes Whitney, Teresa Naval, Elizabeth Quepons, Simrandeep Singh, Steven R Rick, and Lilly Irani. "HCI Tactics for Politics from Below: Meeting the Challenges of Smart Cities". 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: [10.1145/3411764.3445314](https://doi.org/10.1145/3411764.3445314). URL: <https://doi.org/10.1145/3411764.3445314>.
- [569] *Who we are: The beginning of a new era*. 2020. URL: <https://smartafrica.org/who-we-are/>.
- [570] David Gray Widder. "Epistemic Power in AI Ethics Labor: Legitimizing Located Complaints". In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '24. Rio de Janeiro, Brazil: Association for Computing Machinery, 2024, 1295–1304. ISBN: 9798400704505. DOI: [10.1145/3630106.3658973](https://doi.org/10.1145/3630106.3658973). URL: <https://doi.org/10.1145/3630106.3658973>.
- [571] David Gray Widder and Claire Le Goues. "What Is a 'Bug'?" In: *Commun. ACM* 67.11 (Oct. 2024), 32–34. ISSN: 0001-0782. DOI: [10.1145/3662730](https://doi.org/10.1145/3662730). URL: <https://doi.org/10.1145/3662730>.
- [572] David Gray Widder and Dawn Nafus. "Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility". In: *Big Data & Society* 10.1 (2023), p. 20539517231177620. DOI: [10.1177/20539517231177620](https://doi.org/10.1177/20539517231177620). eprint: <https://doi.org/10.1177/20539517231177620>. URL: <https://doi.org/10.1177/20539517231177620>.

- [573] David Gray Widder, Meredith Whittaker, and Sarah Myers West. "Why 'open' AI systems are actually closed, and why this matters". In: *Nature* 635 (2024), pp. 827–833. DOI: [10.1038/s41586-024-08141-1](https://doi.org/10.1038/s41586-024-08141-1). URL: <https://doi.org/10.1038/s41586-024-08141-1>.
- [574] Kyle Wiggers. "Google debuts AI in Google Translate that addresses gender bias". In: *VentureBeat* (2020). URL: <https://venturebeat.com/2020/04/22/google-debuts-ai-in-google-translate-that-addresses-gender-bias>.
- [575] Daricia Wilkinson, Michael Ekstrand, Janet A. Vertesi, and Alexandra Olteanu. *Theories of Change in Responsible AI*. 2023.
- [576] Richmond Y. Wong. "Tactics of Soft Resistance in User Experience Professionals' Values Work". In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW2 (Oct. 2021). DOI: [10.1145/3479499](https://doi.org/10.1145/3479499). URL: <https://doi.org/10.1145/3479499>.
- [577] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. "A Qualitative Exploration of Perceptions of Algorithmic Fairness". In: (2018), 1–14. URL: <https://doi.org/10.1145/3173574.3174230>.
- [578] Lucas Wright, Roxana Mika Muenster, Briana Vecchione, Tianyao Qu, Pika (Senhuang) Cai, Alan Smith, Comm 2450 Student Investigators, Jacob Metcalf, and J. Nathan Matias. "Null Compliance: NYC Local Law 144 and the challenges of algorithm accountability". In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '24. Rio de Janeiro, Brazil: Association for Computing Machinery, 2024, 1701–1713. ISBN: 9798400704505. DOI: [10.1145/3630106.3658998](https://doi.org/10.1145/3630106.3658998). URL: <https://doi.org/10.1145/3630106.3658998>.
- [579] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. "Erudite: Scalable, reproducible, and testable error analysis". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 747–763.
- [580] Yuxi Wu, W. Keith Edwards, and Sauvik Das. "'A Reasonable Thing to Ask For': Towards a Unified Voice in Privacy Collective Action". 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: [10.1145/3491102.3517467](https://doi.org/10.1145/3491102.3517467). URL: <https://doi.org/10.1145/3491102.3517467>.
- [581] Jason Wuertz, Sultan A. Alharthi, William A. Hamilton, Scott Bateman, Carl Gutwin, Anthony Tang, Phoebe O. Toups Dugas, and Jessica Hammer. "A Design Framework for Awareness Cues in Distributed Multiplayer Games". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal QC, Canada: Association for Computing Machinery, 2018, 1–14. ISBN: 9781450356206. DOI: [10.1145/3173574.3173817](https://doi.org/10.1145/3173574.3173817). URL: <https://doi.org/10.1145/3173574.3173817>.
- [582] 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](https://doi.org/10.1145/3491102.3517614). URL: <https://doi.org/10.1145/3491102.3517614>.

- [583] Sijia Xiao, Danaë Metaxa, Joon Sung Park, Karrie Karahalios, and Niloufar Salehi. "Random, messy, funny, raw: Finstas as intimate reconfigurations of social media". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–13.
- [584] Damir Yalalov. *Censorship AI algorithms: How Chinese users cheat them*. 2022. URL: <https://mpost.io/censorship-ai-algorithms-how-chinese-users-cheat-them/>.
- [585] Qian Yang, Richmond Y. Wong, Thomas Gilbert, Margaret D. Hagan, Steven Jackson, Sabine Junginger, and John Zimmerman. "Designing Technology and Policy Simultaneously: Towards A Research Agenda and New Practice". In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI EA '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394222. DOI: 10.1145/3544549.3573827. URL: <https://doi.org/10.1145/3544549.3573827>.
- [586] Yelp. *Yelp's Review Filter Explained*. 2010. URL: <https://blog.yelp.com/2010/03/yelp-review-filter-explained>.
- [587] Leon Yin and Aaron Sankin. *Google Ad Portal Equated "Black Girls" with Porn*. 2020. URL: <https://themarkup.org/google-the-giant/2020/07/23/google-advertising-keywords-black-girls>.
- [588] Meg Young, Sarah Fox, Vinhcent Le, and Oscar J. Romero Jr. *Gear Shift: Driving Change in Public Sector Technology through Community Input*. 2025. URL: <https://datasociety.net/library/gear-shift/>.
- [589] Meg Young, Lassana Magassa, and Batya Friedman. "Toward inclusive tech policy design: a method for underrepresented voices to strengthen tech policy documents". In: *Ethics and Information Technology* 21.2 (2019), pp. 89–103. DOI: 10.1007/s10676-019-09497-z. URL: <https://doi.org/10.1007/s10676-019-09497-z>.
- [590] *YouTube changes restrictions on gay-themed content following outcry*. 2017. URL: <https://www.theguardian.com/music/2017/mar/21/youtube-changes-restrictions-gay-lgbtq-themed-content-tegan-sarah>.
- [591] Zizheng Yu, Emiliano Treré, and Tiziano Bonini. "The emergence of algorithmic solidarity: unveiling mutual aid practices and resistance among Chinese delivery workers". In: *Media International Australia* 183.1 (2022), pp. 107–123. DOI: 10.1177/1329878X221074793. eprint: <https://doi.org/10.1177/1329878X221074793>. URL: <https://doi.org/10.1177/1329878X221074793>.
- [592] Xue Yujie. *Camera Above the Classroom*. 2019. URL: <https://www.sixthtone.com/news/1003759>.
- [593] Omar Zaidan and Chris Callison-Burch. "Crowdsourcing translation: Professional quality from non-professionals". In: *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*. 2011, pp. 1220–1229.
- [594] Sharon Zhou, Mitchell Gordon, Ranjay Krishna, Austin Narcomey, Li F Fei-Fei, and Michael Bernstein. "Hype: A benchmark for human eye perceptual evaluation of generative models". In: *Advances in neural information processing systems* 32 (2019).

- [595] Jonathan Zong. “From Individual Consent to Collective Refusal: Changing Attitudes toward (Mis)Use of Personal Data”. In: *XRDS* 27.2 (2020), 26–29. ISSN: 1528-4972. DOI: [10.1145/3433140](https://doi.org/10.1145/3433140). URL: <https://doi.org/10.1145/3433140>.
- [596] Jonathan Zong and J. Nathan Matias. “Data Refusal from Below: A Framework for Understanding, Evaluating, and Envisioning Refusal as Design”. In: *ACM J. Responsib. Comput.* 1.1 (Mar. 2024). DOI: [10.1145/3630107](https://doi.org/10.1145/3630107). URL: <https://doi.org/10.1145/3630107>.
- [597] James Zou and Londa Schiebinger. “AI can be sexist and racist—it’s time to make it fair”. In: *Nature* 559 (2018), pp. 324–326.