# Ethics down the AI supply chain: playing with power

**David Gray Widder**

CMU-S3D-23-106

July 2023

Software and Societal Systems Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

**Thesis Committee:**
Laura Dabbish, Co-chair
James Herbsleb, Co-chair
Dawn Nafus (Intel Labs)
Jay Aronson

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

**Keywords:** Artificial Intelligence, Ethics, Power, Responsibility, Games, Play

# Abstract

Discussions of ethical issues in Artificial Intelligence have moved from the realm of science fiction and academic debate and into headlines and broad public discussion. Researchers, company leaders, and politicians have ideas about what "Ethical AI" should look like. However, too rarely do these ideas take into account the working realities of the engineers building these systems, to whom ethics work is often relegated.

I conduct qualitative studies of an open source community building a deepfake tool; engineers working across disjointed AI supply chains; practitioners who develop and seek to raise ethical concerns about what they create; and teams discussing AI ethics in a game context. Based on this, **my dissertation shows how organizational norms, incentives, and boundaries limit software creators' sense of responsibility and agency over the downstream impact of what they create, and examines the possibilities and shortfalls of play as a way to expand these limits**. I find that limited conceptions of and authority in their roles, limited visibility into downstream uses, time pressures, licensing constructs, and other reasons often mean engineers do not feel responsible for–or power to do–ethics work. I conclude by discussing how future tech ethics research, practice, education, and policy can better consider the sense of responsibility and power of those asked to do ethics work in technology.

# Acknowledgments

Thank you to my parents, Julie Martinson Widder and John Michael Widder, and sister, Elizabeth Rose Widder, for being my first teachers and fellow curious learners.

Thank you to my thesis committee: my thesis co-chairs Laura Dabbish and Jim Herbsleb, as well as Dawn Nafus and Jay Aronson for their diverse perspectives and support through the research and writing process. I have been pushed to grow as a scholar in unique ways from each of you, and I am dearly grateful for this.

Thank you also to the many other mentors for their support before and during my degree: Claire Le Goues, Joshua Sunshine, Jonathan Aldrich, Nick Frollini, Bogdan Vasilescu, Alexandra Holloway, Scott Davidoff, Krys Blackwood, John Sherry, and Nik Martelaro. Thank you to Christian Kästner for keeping his door open for conversation, even difficult ones, and his chocolate bowl stocked.

Thank you to the many people which made my graduate student experience more bearable: my partner Stephanie Gray, my friends Leo Chen, Maggie Oates, Asher Trockman, Clair Hopper, George Barrow, Amanda Bertsch, Sireesh Gururaja, Clara Na, Robert Petersen, Es Braziel, Maria Ryabova, Leif Hancox-Li, Franky Spektor, Cella Sum, Krystal Jackson, Jessica Colnago, Kalil Anderson Garrett, Lucio Dery, Bryce Sprauer, Marat Valiev, Dasha Pruss, Jane Hsieh, Roykrong Sukkerd, Jens Meinicke, Chu-Pan Wong, Bonnie Fan, and Kyle Liang; my office mates Momin Malik, Hongbo Fang and Yining She; and my office neighbors Mike Skirpan and Chris Bogart. Thank you to the many other PhD students, including Zack Coker, Victoria Dean, Lisa Egede, Jordan Taylor, and Huilian Sophie Qiu; and those convening from across the world for our Critical AI Studies Reading Group, for the camaraderie. Thank you to my therapist, Lynn Allen.

Thank you to the five undergraduate research mentees whose ideas were instrumental to my own development, and whose work made possible some of the research presented here: Courtney Miller, Jamie Rosas-Smith, Hana Frluckaj, Lyric Sampson, and Derrick Zhen. I am so proud of you all, have learned so much from you all, and am excited to see where you go. Thank you to those leading the REUSE undergraduate research program in which I have been a grateful mentee and proud mentor, including Josh and Claire, and also Charlie Garrod, Michael Hilton, and Samantha Mundrich.

Thank you to the many who have helped develop my thinking by providing feedback on ideas presented here and elsewhere: Karly Burch, Henry Fraser, Michael Madaio, Lama Nachman, Richard Beckwith, Suzanne Thomas, Richmond Wong, Mary Shaw, Emma Strubell, Maarten Sap, Seda Gürses, Danielle Wenner, Elizabeth Anne Watkins, Valerie Pilloud, Blair Attard-Frost, Ken Holstein, Sarah Fox, Aspen Omapang, Amy Ko, Sarah Myers West, and Meredith Whittaker. I am grateful to be in community with Noah Theriault, Jordan Taylor, Lisa Egede, Reuben Aronson, Abhijat Biswas, Em Cariglino, ocean, Coraline Ada Ehmke, Catherine Taipe, Renée Nikolov, Kiyn Chin, Os Keyes, and Lauren Herckis. Carol Frieze opened many doors. To all of those who participated in this research, thank you for your

reflection, frankness, and valuable time.

Thank you to the administrative staff whose work makes everything here work: Victoria Poprocky, Connie Herold, Nancy Beatty, Tiffany Todd, Dabney Schlea, Nancy Beatty, Catherine Copetas, Paul Emanuel Bowes, Tom Pope, and more whose work is invisible to me though no less valuable. Thank you also to Tay, and the other custodial and food service staff whose work makes the present work possible yet whom I did not meet or whose names I didn't learn. May Carnegie Mellon negotiate with your union in good faith, and not build tech to surveil you while you work.

I would not be here without my undergraduate and high school teachers. At the University of Oregon, thank you to Stephen Fickas, for being the first to invite me to work on a research project; Joseph Fracchia, for helping me sharpen my written and oral argumentation through a love of historical political thought; and to my advisors Helen Southworth and Kathleen Freeman Hennessy for guidance and encouragement as I struggled through the beginning of undergrad and looked toward my future. In Singapore, thank you firstly to Helen Leeming, for getting me almost kicked out of high school (deservedly!), and then her endless work and care to help me thrive. Thank you to Alex Varghese, for patiently teaching me computer science fundamentals. Thank you to Qamaruzzaman bin Amir, for role modeling a curiosity about the universe I cherish, through his endless hours helping me figure out everything from physics formulas to metaphysics. For the extra motivation when writing my PhD applications, I am grateful to my eighth grade math teacher, who replied "hmm, why don't you set your sights a little lower?" when I confided in him that I hoped to apply to study computer science at Carnegie Mellon one day.

I hope that the primary impact of my past six years is not what is discussed in this document. In coalition with other students, faculty and staff, we've fought for us graduate students to have access to healthcare we can afford, have workspaces with daylight, and to be included meaningfully in organizational governance. We have stood behind custodial union stewards as they negotiate with CMU admin [8], demand that this institution stop hosting recruitment events for Immigration and Customs Enforcement and its contractors especially while this agency separates kids from their families **#NoTechForICE!** [60], and fought for anti-racist changes to School of Computer Science policies on recruitment, admission, hiring, and promotion [229]. We have protested the technology Carnegie Mellon University forces upon its community: questioning this institution's partnership with the Pittsburgh Police to develop and deploy racist predictive policing technology [1] as well as campaigning to regulate the latter's use of facial recognition [72], my own department's non-consensual installation of networked sensing devices in the offices and common areas of our new building [39, 110]. Not all of these struggles have been successful, and there is much I wish I did differently, but I believe this work has enabled new coalitions and discursive possibilities to form on campus. Many of my fellow organizers are still employed at CMU—so I do not name them—but they are why I began to ask more critical questions about technology and power relations, and so this research is inextricably, imperfectly and gratefully shaped by these people.

# Contents

# Chapter 1

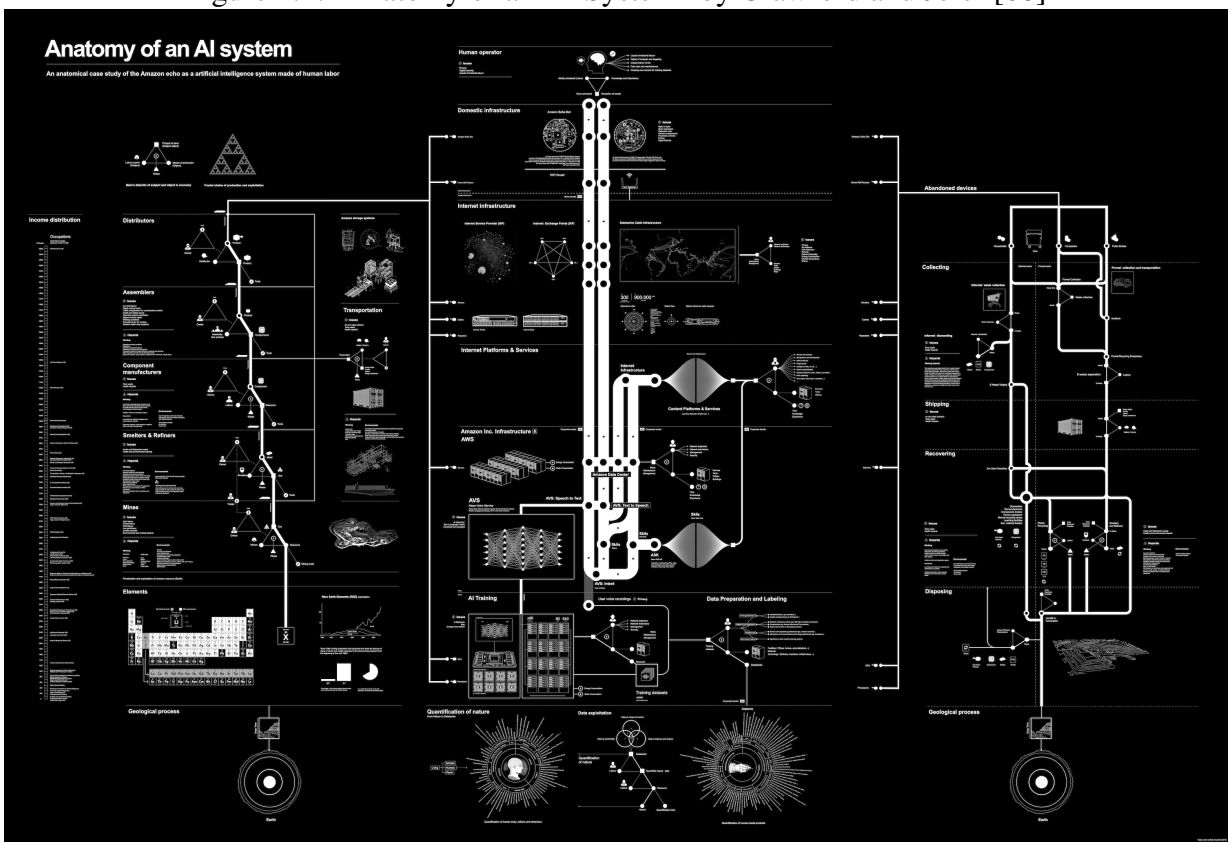
# Introduction and Related Work

## 1.1 Dissertation Overview and Thesis Statement

So-called 'smart' devices are an example of an Artificial Intelligence (AI) technology becoming lodged in almost every aspect of human life, but their creation, design and use have implications most are unaware of. A diagram entitled "Anatomy of an AI System" by Crawford and Joler [68] (shown in Figure 1.1) artistically demonstrates all that is needed to create, operate and dispose of an Amazon Echo "smart home" device. This includes workers mining rare earth minerals for its components, network engineers in data centers where its cloud processing occurs, crowd workers labeling data to train its voice recognition system, and the software engineers building this system. This outlines the vast extent of possible labor, data, and natural resource issues we might wish to scrutinize when discussing the ethics of just one AI-enabled product.

The biggest tech companies now have AI ethics guidelines, often converging on design principles seeking to ensure AI systems are "Fair" or "Transparent" [137]. Instruments like checklists seek to translate these guidelines into work for software engineers building AI systems [165]. However, these checklists and toolkits often frame AI ethics work as *technical* design considerations to be examined by the *individual* practitioners, often alone, and without mechanisms to help build power, such as by acting collectively, to enact changes [277]. Other work shows

Figure 1.1: ‘Anatomy of an AI System” by Crawford and Joler [68]




how corporate AI ethics initiatives seek to individualize risk in a way that “puts the onus on individuals to take on work” [14], in lieu of substantive investment in or regulation of ethics. Beyond those doing the work, AI ethics guidelines are often limited in scope to scrutinizing and enabling the *design* of AI systems, rather than scrutinizing the business uses these systems are put to [108], and neither technical changes [142] nor convergence around ethical principles [185] guarantee ethical outcomes. When we attempt to locate software practitioners instructed to build more ethical systems on Crawford and Joler’s vast web, we begin to see how incomplete this narrowly scoped framing of AI ethics may be. So as we see, there are limitations in how “Ethical AI” is framed, and how work to implement “Ethical AI” is given to and received by the engineers building it.

Nonetheless, this dissertation focuses primarily on these engineers. Why? Well, as we will

see, while these engineers often only have narrow authority and expertise over the technical implementation of systems, this provides some latitude for ethical action. Their intimate knowledge of and exposure to the technical limitations of their systems provides a basis for some to think and talk about negative ethical impacts which may result. Then, there is the reality that, while there are myriad ethical concerns which better engineering cannot fix, ethics is often left to engineers. They are the ones most acutely caught between lofty ethics rhetoric and the working realities of building large scale systems. Therefore, as those caught in the bind between flawed demands for ethics and their own limited scope of knowing and acting, I hope to demonstrate that a rich understanding of their technical practice and working realities can be a fruitful basis for critical reflection on tech ethics.

My work seeks to understand how these engineers building software systems think about the ethical impact of what they create, and what they want to and are able to do about it, especially under the constrained notions of ethics as I describe above. In particular, **this dissertation examines how organizational norms, incentives, and boundaries limit software creators' sense of responsibility and agency over the downstream impact of what they create, and examines the possibilities and shortfalls of *play* as a way to expand these limits.**

My dissertation proceeds as follows. In the rest of this chapter, I review past work on concepts which situate the rest of this dissertation: "Artificial" "Intelligence", what "ethical AI" and "doing ethics" often mean, and work which informs my consideration of power and agency in technology organizations, and I leave more detailed discussion of other prior work encapsulated in the chapter where it is most relevant. After this, in Chapter 2, I show how open source norms lead the creators of a Deepfake tool to disavow ethical responsibility for how it is used downstream. In Chapter 3, I generalize this to thinking about AI supply chains—assemblages of existing modules used to build contemporary AI systems—to show how organizational boundaries lead AI creators rarely feeling able to control or responsible for harms in how their systems are used. Then, in Chapter 4, I examine self-identified ethical concerns that software engineers

raise in their work, and show the barriers they face as they seek to resolve these concerns. In a final study in Chapter 5, I examine discursive closure in how teams discuss AI ethics, and the possibility of using a game to reveal and question this closure. In Chapter 6, I discuss broader implications across these pieces of work. In particular, I discuss how my examinations of power and the supply chain concept can be applied in support of future research, practice, education, and policy.

## 1.2 What is Artificial? What Intelligence?

Many of the people who participated in the research I present in this dissertation identify themselves as working on "Artificial Intelligence" ("AI"), a project which I describe here. I use "AI" as an emic term, recognizing its significance to those within its epistemic community [9], while seeking to maintain critical distance from the epistemic claims and ideals embedded within it [37].

One place to locate the beginning of an academic project to create "Artificial Intelligence" could be the 1956 "Dartmouth workshop". Here, a group of men including names like Marvin Minsky, Herbert Simon, and Allen Newell convened on the "the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" [193], in hopes of allowing "machines [to] use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves". Their quest for artificial intelligence, then, was to build a machine with human intelligence and abilities.

There are many technical approaches to creating AI, including symbolic approaches attempting to encode knowledge about the world into a set of rules in an approach embraced by Minsky, Simon and Newell; and today's dominant machine learning approaches which use statistical techniques to infer rules from data about the world [97, 172]. Within machine learning, there are supervised approaches in which humans sort data into preexisting categories (called labels)

to train a computer program to learn how they are different, unsupervised approaches in which unlabeled data is given to a computer so that it may infer its own categories based on similarities between data points, and reinforcement learning approaches in which useful intermediate behavior is learned by evaluating how well it achieves a specified overall goal [248]. I use words like "train" and "learn" skeptically: they are emic terms of art within AI and in wider use, but are also examples of "intentional" vocabulary (*i.e.*, ascribing intention to machines) [283]. This constitutes a discursive practice of gesturing towards vague but intuitively understood human traits, which are nonetheless formalized in a precise and closed manner, enabling the tendency to conflate representations with things themselves and hampering attempts to build more critical approaches to AI [9], and perhaps foreshadowing AI hype in our current era.

In 2018, AI researcher Zachary Lipton warned that "It's getting harder and harder to distinguish what's a real advance and what is snake oil", and that "AI hype" may lead people to place too much trust AI in high stakes cases like "autonomous" vehicles or clinical diagnoses [176]. Employee researchers at Microsoft claimed that the GPT-4 released by their business partners at OpenAI displays "sparks" of "Artificial General Intelligence" (AGI) [45], in a non-peer-reviewed paper that a non-Microsoft scientist criticized as a "press release ... masquerading as science" [173]. The Microsoft employees supported their claims of AGI by pointing to GPT-4's ability to draw successively sophisticated clip art unicorns, and correctly answer legal and medical licensing exam questions, among other tasks [45]. However, they concede that neither "AGI" nor even "Intelligence" are well defined, and also that they "do not have access to the full details of its vast training data [and thus] we have to assume that it has potentially seen every existing benchmark" [45].

Pulling on the "vast training data" thread helps unravel hype-ridden yarns of "Artificial Intelligence". While those claiming to see sparks of Artificial General Intelligence in GPT-4 nonetheless concede that intelligence is ill-defined [45], and OpenAI cites a "competitive landscape" and "safety implications" in refusing to release details about GPT-4's dataset [200], more critical

work shows how AI is not itself intelligent, but instead dependent on distilled human labor and data. Among AI practitioners, work scrutinizing curating datasets is seen as low-status work relative to developing new models [225], so this work is outsourced to vast arrays of humans abstracted and abused behind platforms such as Amazon's "Mechanical Turk" [25, 128], or other poorly-paid outsourcing arrangements [5]. As we saw earlier in the introduction, Crawford and Joler reconstruct the vast network of human labor and data needed to allow a home voice assistant to function [68]. Others warn that training models using vast and uncurated datasets scraped from the internet can make it difficult to scrutinize for and remove bias, and also cause environmental damage [30]. In summary, AI practitioners may therefore have little awareness of what is in their dataset, whether due to its size, the low status nature of data work, a refusal on the part of companies to release their data, or because outsourcing arrangements lead to their ignorance of the conditions in which it is produced.

This is not to undersell the very real technical capabilities of AI. It is not all "snake oil", AI *can* be used to do powerful things, which is why we must scrutinize these systems, how they are created, and what they are used for. To illustrate what AI is capable of without enumerating every use, here are two. Firstly, models can recognize faces in images, and identify the same face in different images. However, work has shown how error rates when performing this task can differ for faces of different genders and skin types [46], which, when AI systems are deployed in high-stakes settings, may disproportionately deny minoritized user groups access to important services. Large language models can also now produce grammatical and seemingly plausible text in response to human prompts, such as in the widely-hyped ChatGPT [45]. However, ChatGPT also presents incorrect or harmful answers in the same convincing tone it presents correct ones, leading to risks when it is deployed in high stakes settings (*i.e.*, eating disorder counseling [278]). More generally, large language models can also absorb bias from their training data, and reflects this bias in their output [30].

## 1.3 "Ethical AI" and Doing Ethics

Workers in technology companies have raised ethical concerns with the tech their companies are building. This year, "godfather" of AI Geoffrey Hinton resigned from Google "so he can freely speak out about the risks of A.I", [180], saying that "this stuff could actually get smarter than people". Only a few years prior, other Google employees were forced out. Meredith Whittaker left Google, after retaliation [63] after her organizing to protest the company's handling of sexual misconduct and against company plans to use AI to help the US military target military drone strikes [131]. Timnit Gebru [181] was forced out after being told to retract her paper ([30]) detailing the risks Large Language Models pose to the environment and in perpetuating bias.

Yet at the same time, "Ethical AI" guidelines abound. One analysis of 84 such guidelines suggest a convergence around principles like "Transparency" to enable system interpretability, or "Fairness" to enable the "prevention, monitoring or mitigation of unwanted bias" [137]. There are myriad checklists [165] and toolkits [277] to ensure AI is fair, processes to design transparent models [87], and ways to document model and dataset biases and limitations [98, 184].

There is also a variety of academic research on techniques to enable attributes like transparency and fairness in AI systems. For example, there are various computational approaches to examine and improve model fairness [80], and analyses of how different mathematical definitions of fairness have different implications in high stakes settings [34]. There is also research into techniques to visualize the inner workings of machine learning models [285].

Given this research, and given that "Ethical AI" guidelines are in wide use—including at Google—why are Whittaker, Gebru and Hinton no longer at Google after high profile departures? Scholars have argued that convergence around principles will not guarantee Ethical AI, as among other problems, they hide deep normative disagreement [185], and that these principles focus scrutiny onto the narrower question of *how* these systems should be designed and built, thereby eliding other critique such as that of business uses [108], or *whether* we should build these systems for a given purpose at all. Those seeking to affect ethical change within their orga-

nizations face limits in doing so under tech industry logics [179]. Analysis of AI fairness toolkits show how they frame AI ethics work to be "technical work for individual technical practitioners", uncovering their unwritten and dubious assumption that practitioners can solve AI ethics concerns themselves, and do so with narrow technical fixes. Others criticize design-stage "Ethical AI" interventions as not preventing harms which occur after something is deployed, such as through intentional misuse [96]. The limitations of these principles have also been illustrated with pointed satire, one example describing a system used to turn elderly people into milkshakes in a Fair, Accountable and Transparent way [142]. Even the creators of widely-cited AI fairness checklists warn that practitioners may engage with such checklists through a "minimal, superficial completion of items" [165].

Ethics is a subfield of philosophy which seeks to examine what is right and wrong [207], and within Western philosophy, there are three major theories on how moral conduct should be guided or evaluated. Deontological stances, closely associated with the Enlightenment-era philosopher Immanuel Kant, argue that an action holds moral worth when done from a place of duty by following moral rules [210, 232]. Teleological (consequentialist) stances emphasize outcomes: that the ethics of an action ought to be evaluated on its consequences, often on the basis of utilitarianism, which seeks to ensure the most good for the most number of people [210]. Finally, virtue ethics stances emphasize the goodness of one's character, over how one acts or follows ethical rules or the consequences of one's actions [210]. Non-Western perspectives on AI ethics include those based on the philosophy of Ubuntu common in parts of Africa and including values such as interdependence or solidarity [18], and Buddhist perspectives integrating karma [159].

This dissertation does not adopt one of these stances, nor seek to arrive at a particular definition of ethical behavior, but instead examines how practitioners' "tacit definitions" of ethics are revealed as "part of [their] in-situ processes" [201]: seeing how people building software understand their own obligations, and the way they attempt to fulfill these obligations. Within technology companies, this often but not always involves using these guidelines, toolkits and

checklists, which can comprise much of what it means to be "doing ethics" work. This thesis examines this kind of work, but also how practitioners' sense of obligation may not be addressed within the limited scope these ethics processes offer, and outside of company contexts.

While I do not adopt a particular view of what is ethical, Lucy Suchman's notion of located accountability suggests preconditions from which creating ethical technology might emerge. She writes that responsible technology development must be "a boundary-crossing activity, taking place through the deliberate creation of situations that allow for the meeting of different partial knowledges" [247]. In this sense, allowing and encouraging those involved in a system's creation and use to stay in relation to one another, and speak from their own partial perspectives about their ideals and concerns, would be a precondition for us to agree on what "ethical AI" would entail.

## 1.4 Power and Agency in Organizations, Supply Chains, and when Discussing Ethics

Many have made calls to center power in tech ethics work. This includes calls to examine power asymmetries in concrete contexts [38], the role of corporate power in subordinating ethics concerns [96], the role of corporate power in driving academic AI ethics discourse [196], and to recognize the power of "who gets to make decisions?" in the building and deployment of technology [155].

The Oxford English Dictionary defines power as the "capacity to direct or influence the behaviour of others" [213], and the related concept *agency* as the "ability or capacity to act or exert power" [212]. However, a plethora of writing exists about the manner in which these capacities are obtained, and about their effects [33, 91, 92, 153, 228]. In this dissertation, I draw upon prior work to discuss power and agency in three main contexts: 1) frameworks of power from organization science to examine how power operates on workers within companies or other institutions,

2) when inter-organizational divisions fracture power over ethics in wider "supply chains" of AI, and 3) when discursive closure shapes what practitioners feel able to raise in discussions of tech ethics, often implicitly. In this dissertation, I do not adopt a single theory of power or agency, but draw on several pieces of prior work most relevant to each component study.

### 1.4.1 Faces of power in organization science

I draw on notions of power from the field of organization science to understand how authority, hierarchy and incentives within companies affect practitioners' ability to raise and resolve their own ethical concerns. In Chapter 4, I discuss the faces of power within organizations that leave software engineers who develop ethical concerns with their work often unable to raise them or ensure they are address, drawing primarily Fleming and Spicer's framework of power [91] because it deals most directly with the mechanisms of how each face of power works. When relevant in this chapter, I secondarily draw on the forms of resistance that Lawrence and Buchanan discuss [153], as well as notions of disempowerment detailed by Berti and Simpson [33].

Fleming and Spicer's review and framework of power within organization science [91] includes four "faces" of power. Two are "episodic" in that they occur in identifiable instances: *coercion*, directly exercised where one is "simply told what to do 'or else"', and *manipulation*, where actors seek to limit what is discussed to acceptable boundaries [91]. Two faces are "systemic", where power is "congealed into more enduring institutional structures": *domination*, where actors establish influence by constructing hegemonic ideologies; and *subjectification*, which seeks to influence an actors' sense of self, emotions and identity [91].

Berti and Simpson respond to the lack of full agency in how actors are able to respond to organizational "paradoxes" in which there are contradictory and interdependent demands, by surveying management literature to create a framework of disempowerment [33]. They categorize experiences of disempowerment under the four faces of power that Fleming and Spicer propose, and demonstrate how institutional paradoxes like double-binds, catch-22 situations, and

doublethink are required of employees, helpful in trying to understand contexts where workers may be told to "do ethics" without being given power to make requisite changes.

Lawrence and Buchanan propose a different framework for understanding the relationship between power and institutions. The terms used in their framework sometimes overlap with those in Fleming and Spicer's framework yet often mean different things, but Lawrence and Buchanan's particular attention to the forms of agency employees have to resist institutional power (which they term"institutional agency") proves useful as I seek to understand what may give workers power to advocate for ethics. For example, they write that some "employees [who are] professionally mobile (based on skills or family connections)" would be better able to resist institutional control.

### 1.4.2 Locating power in chains of partial knowledges

Thinking back to Crawford and Joler's diagram, we also must ask how power and agency is fractured across the organizational divisions which enable different parts of software systems. To do this, I look to the work of Lucy Suchman, and her notion of *located accountability*. Drawing on Donna Haraway's writings on Feminist situated knowledges [115], Suchman's located accountability casts responsible technology development as "as entry into the networks of working relations" [247], which involve locating those who have power over different parts of its production, and recognizing the "contests and alliances" between them.

Lucy Suchman discusses the power of knowledge and those who produce it, in a call to move away from "objective knowledge as a single, asituated, master perspective that bases its claims to objectivity in the *closure* of controversy" (emphasis added, foreshadowing discussion of closure in the next subsection), and toward a feminist conception of objectivity involving a meeting of "partial perspectives" through which knowledge is produced in ongoing debate among all parties, who are then responsible for what emerges. Suchman quotes Donna Haraway [115], who writes that we must "translate knowledges among [....] power differentiated [...] communities", and

Suchman argues that doing so would entail "acknowledging and accepting the limited power of any actors or artifacts to control technology production/use", each having limits to what they know and are able to act on.

A view of power informed by Suchman, thus, would involve asking who is in relation with whom, and what is rendered possible as a result of these relations: between those who may have knowledge of, and power over, different parts of sites of production and use of technology. This view of power is foreshadowed in Chapter 2 through mention of the metaphor of the "supply chain", and then used extensively in the following Chapter 3 which deals with supply chains in their technical and cultural effects directly.

### 1.4.3 Power in discussions of tech ethics

Organizational norms and processes may also affect what differently situated actors feel is within their power to do and say when discussing contested topics, notably tech ethics. I primarily draw upon two pieces of existing scholarship to explore this, to explore how people speak in the presence of power, and to examine how organizational dynamics can enact discursive closure.

In his book "Domination and the Arts of Resistance", James Scott drew from his fieldwork to show how people speak differently depending on power differentials between them and their audience. He wrote about how less powerful subjects use "public transcripts" when speaking in the presence of those with power over them, but will persistently use "hidden transcripts" to challenge power when speaking "offstage ... outside the intimidating gaze of power" [228]. Scott emphasizes continuity between these two "stages", and that "rumors, gossip, folktales, songs, gestures, jokes" are the places where people may demonstrate dissent more freely while "hiding behind anonymity or behind innocuous understandings" [228]. In Chapter 3, I discuss how offstage talk and actions may serve as a way to connect dislocated modules in supply chains, and in Chapter 5, I examine whether games might provide an opportunity to create offstage contexts for this to occur.

The limited scope of AI ethics principles [108] may be seen as an instance of *closure*. Closure is a "rhetorical process through which relevant social groups perceive their problems with an artifact to be solved or closed" [125]. In their analysis of sustainability governance standards within organization, Christensen *et al.* argued that business contexts often push towards closure, which they define as "the termination of reflection and debate about what sustainability means or could mean for organizations and society". To guard against this closure which can "squeeze out open debate and deliberation", they argue that a deliberate "license to critique" must be deliberately enabled via intervention to enable discursive openness [57]. Recognizing the power dynamics that may lead employees to avoid debate rather than resist this closure, Christensen *et al.* write that this openness must involve "empowering" participants, and that "articulating ideals and ambitions out loud is an essential sense-making mechanism that needs to be stimulated and protected by management" [57]. Thinking with Scott and Christensen *et al.* together in the context of responsible technology development, those who codify Responsible AI standards [137], and those who operationalize them within a particular context, have power: to define what is easier to say, and that which may be too risky and thus sayable only "offstage" or though "innocuous" coded dissent. We may expect power to influence the conditions under which workers are willing to raise concerns about the ethical implications of their systems. I draw on this notion of power through closure in Chapter 5, where I examine how teams of engineers and activists—with their existing relationships, hierarchy and norms—discuss AI Ethics, and examine whether a game format can enable this "innocuous" context for dissent.

# Chapter 2

# Open Source Norms and Contributors' Ethical Responsibility

Work in this chapter was originally peer-reviewed and published at the 2022 *ACM Conference on Fairness, Accountability, and Transparency*, with coauthors Dawn Nafus, Laura Dabbish, and James Herbsleb [268].

## 2.1 Introduction and Related Work

Discourses of "Ethical AI" have largely focused on issues that arise in software produced by private companies. The drafters of the frequently cited "Montréal Declaration for a Responsible Development of Artificial Intelligence" [7] asked if we must "fight against the concentration of power and wealth in the hands of a small number of AI companies" in early deliberative discussions [6, 108], and indeed we should. However, an important perspective and site of AI practice is largely missing from "Ethical AI" discourse: Free and Open Source[1] developers creating AI software, who have unique limitations on and possibilities for ethical action. Open source AI development is significant: for example, two of the most popular AI libraries are open source:

[1] *Open Source* eschews *Free* software's ideology; we use "open source" here. See Sec. 2.4.1.

SciKit learn, and TensorFlow (after being open sourced by Google), along with myriad end-user AI projects. While harm does originate from a concentration of AI power in companies [265], we show that significant and understudied harms originate from differing practices of transparency and accountability in the open source community.

A 2019 systematic analysis of 84 "Ethical AI" guidelines [137] found that most guidelines are produced by private companies (22.6%) or governments (21.4%) often seeking to regulate AI from private companies. Abstract "Ethical AI" principles (*e.g.*,, "transparency", "interpretability") are used with differing underlying meanings, and apparent convergence may be superficial [137, 161]. Systems may adhere to such principles while still being patently unethical [142], and convergence on principles risks obscuring political and normative disagreements [185], or focuses "Ethical AI" scrutiny on AI design rather than the business uses it enables [108]. Even critical discourse often focuses exclusively on the private sector: one study found that "principles alone cannot guarantee Ethical AI", but stated in their introduction: "AI is largely developed by the private sector" [185].

When design, policy and tooling interventions to encourage "Ethical AI" are built with private companies in mind, they risk being ill-suited for an open source context. For example, facing employee rebellion, Google decided to stop providing the US military with AI which could be used to improve drone strike targeting [261]. This decision was undoubtedly *politically* fraught, but enacting it was *procedurally* easy: the company exercised its legally available and enforceable right to not renew a contract. However, open source supply chains are messy: code is reused, and projects are copied and adapted (forked) [282], and it is difficult to track, constrain, or assign accountability for downstream uses. Conventional notions of accountability rely on stable entity to hold accountable, whereas open source membership can be unstable [198], and some even contribute anonymously [71].

Crucially, these structural challenges have cultural underpinnings. [62] The founders of the influential Free Software movement advocate for "Freedom 0"– the right of anyone to reuse code,

for any purpose [238], encoded into legally binding licenses – and decry attempts to abridge this freedom even in service of other ethical ends [237]. Similarly, Transparency is often held as a near-universal principle in "Ethical AI" guidelines [137], but others reason how openness may not be universally desirable, giving autonomous weapons development as one example [41].

Studies to help AI practitioners improve fairness [67, 122], such as checklists to solve organizational challenges [165], are often based on the needs of AI practitioners in private companies, but some studies also focus on the needs of public sector [258] or academic institutions [201]. These results expose the role of organizational structures in AI Ethics practice, structures which look very different in open source. On the other hand, incentives in private organizations can hinder "Ethical AI", where developers work in "an environment which constantly pressures them to cut costs, increase profit and deliver higher quality [systems]" [256], and "face pressure from management to make decisions that prioritize company interests"[170, 185], and companies compete in a wider market structure which can hinder "Ethical AI" work [179]. Alongside the possible challenges for "Ethical AI" in open source we discuss in this chapter, we also see a cause for optimism: unconstrained by these forces, experimentation may be more possible in open source communities to offer new ideas to solve ethical challenges unsolved in company contexts, or provide space to challenge assumptions made in private companies' "Ethical AI" endeavors.

To begin reconciling conflicts between norms in open source communities and prevailing assumptions in "Ethical AI" discourse, we ask: **How do members of an AI-enabled open source Deepfake project reason about the ethics of their work?** To answer this, we conduct an interview study in an open source community which builds software to create "Deepfakes": videos which replace the likeness of one person with another [143]. The community celebrates artistic and educational uses they see as ethical, and explicitly takes a position against and actions to discourage uses they believe are unethical, such as non-consensual or child pornography and fake news. In our study, we uncover normative, structural, and technical barriers to the commu-

nity achieving their stated ethical views, and situate these barriers within the dominant private-company-focused "Ethical AI" discourse and political tensions in the open source and wider tech worker communities. In the additional appendix, we outline ideas that open source communities and platforms may want to experiment with, which researchers may also be interested in evaluating and studying further.

## 2.2 Methods and Setting

### 2.2.1 Setting

We set our study in an open source Deepfake creation tool, an AI technology with contested ethical issues [143], positioning it as an extreme case study [280] where ethical reasoning and its situated relationship to other cultural frames may be especially apparent. A 2019 study found that 96% of online Deepfakes are non-consensual pornography, 99% of which depict women celebrities [11]. Scholars write that political Deepfakes operates similarly to non-consensual Deepfake pornography to silence critical speech, and that victims of the latter experience anxiety, illness and job loss [167]. Other scholars explore how Deepfake distribution enforces gendered disparities in visual information [260], and find that more attention in public discourse is given to viewers of Deeepfake disinformation than do the women depicted in Deepfake porn [103]. One study analyzed Reddit and GitHub posts and found a tension between between moderation practices and open source ethos, recommending future work beyond identifying or regulating Deepfakes to understanding their antecedent code and programmers which enable their creation. [273]. We do not seek to define or evaluate ethical behavior, which others studying AI practitioner's views on ethics (*i.e.*, [201]) recognize as an entire branch of Philosophy, with divergent proposed approaches in AI [19, 24, 108, 188]. Instead, we examine "how AI practitioners understand the ethical landscape and their own role within it" [201], including "procedures, decisions [... and resulting] related responsibilities" [201], and examine how their perspectives do or do

not fit with prevailing AI ethics discourse.

The first widely-available face swapping algorithm was posted by an anonymous user in a Deepfake-focused sub-Reddit [11, 143, 273], which has since been banned for violating the site's more recent "policy against involuntary pornography". This algorithm became the basis of many open source projects; we approached the project of our study because of its unique willingness to engage in questions of ethics as indicated by its public ethical stand. The project's original leader copied this algorithm from Reddit to a repository on the social coding platform GitHub [70], which new leaders use to track code changes and bugs, and host usage instructions and a contributor guide. Current leaders rewrote the codebase and applied a GNU General Public License (for implications of this, see Sec. 2.3.1). The GitHub project page prominently features a statement written by the project's leaders to explain the benefits of releasing the software publicly, such as enabling AI learning, political commentary, and artistic uses, while acknowledging and claiming a refusal to support non-consensual, inappropriate, illegal, unethical, or questionable uses. The GitHub project page directs support requests to two other platforms: a Discord chat server and a self-hosted online message board. On all platforms, there is an expressed "Safe For Work" policy, for example, one is posted in the "Welcome" section of the Discord chat server, which states that even discussing NSFW content will result in an immediate ban without further warning. These platforms provide space for the 500+ users who are often online at once to seek and provide technical support, share Deepfakes they have created, and discuss broader Deepfakes and AI issues. The leaders are informally designated, often being invited to join private channels and given administrative privileges by existing leadership after contributing to the project codebase, or by creating high quality Deepfake content. These leaders use these channels to discuss development and moderation decisions, which they have broad discretion to make independently. This project is not corporate affiliated, but accepts donations. Users often used humorous display names, but established users often knew each other's real names. The first author observed a generally collaborative and polite tone in these venues.

### 2.2.2 Recruitment, Participants, and Data Analysis

The first author approached project leaders who gave permission to recruit in their community and collaboratively crafted a recruitment message which a leader shared in the project's chat server, resulting in eleven completed interviews. All self-identified as male, and were mostly from the United States and Europe, resembling open source generally, but the modal age range was 35-44, somewhat older than open source generally [99]. For confidentiality, we do not discuss individual demographics. Participants had a median of 7 years of programming experience, 2 years of AI experience, and 2 years working with the project in roles ranging from developing and testing the project code, supporting users, content moderation in communication channels, and both hobbyist and professional users of the project. The first author conducted semi-structured interviews in a one-on-one setting due to the possibly sensitive nature of topics discussed [264]. Most interviews were conducted via a teleconferencing call and lasted 30-93 minutes, with most lasting about an hour. In two cases, chat interviews (*i.e.*, [240]) were used for accessibility reasons.

We adopt an *interpretivist* epistemological paradigm [160]: the framings presented below emerge from the intersubjectivity between researcher and participant, and cultural frames they do and do not share. We also observed chat room discussion and work interactions on GitHub, but we acknowledge that self-reports from our primary interview method may hold limited value in explaining behavior and attitudes in actual context [135], and caution that there are meaningful differences between open source communities that limit the ability to generalize these findings to the exceptionally organizationally and politically diverse landscape that is open source. We note that the male dominance in this community and Deepfake production communities generally contrast sharply with the vast majority of online Deepfakes which non-consensually depict women in pornography [11], and past work which we discuss in Section 2.2.1 has discussed the gendered politics of Deepfakes, and future work using feminist analytical frames could unpack gender dynamics of how exclusion plays a role in the choices that open source communities see

as available to them.

Data was analyzed in an iterative process including a descriptive memo after each interview, and a running analytic memo as a reflexive history of the first author's understanding of emergent themes [243], and weekly discussions among the research team to discuss commonalities and contrasts between interviews. After data collection, all interviews were transcribed, and then the first author examined possible relationships between themes in this analytic memo, iteratively going back to the data to test out these possible structures, before settling on an inductive hierarchical coding frame [163, 183, 249]. This was then used to code the entire dataset. During this coding process, our understanding of the data deepened and new codes arose to capture new themes or provide greater specificity, in which case an open card sort was used [219] to identify sub-codes, after which the dataset was re-coded.

## 2.3 Findings

### 2.3.1 Responses to Ethical Issues

Participant's perceptions about what they could and couldn't do about Deepfake misuse was shaped by open source licensing, discourses about progress and the neutrality of tools, and by setting community norms of acceptable use.

#### Open Source Licensing as a Frame for Ethics

We saw that the open source license of this project is highly relevant to participants' ways of understanding their responsibilities, and therefore their responses to the problem. It is both a legal set of constraints that sets out what developers can and cannot do to prevent uses they view as unethical, and a normative one that frames broader cultural values beyond what the license requires (see also [62, 140]). Leaders lamented that, as they saw things, the open source status of their project (a choice they made) prohibits them from controlling downstream uses. A leader

remarked: *"We've got very limited control. [...] We can't prevent people from getting access to a software using it. [...] Part of being open source Free Software is that you are free to use it. There are no restrictions on it. And we can't do anything about that." (2)*

Even if the leaders wanted to choose a more restrictive license for their project, the leaders' prior choice of a GPL license led contributors to view applying a more restrictive license as impossible at this stage: *"Anything that touches GPL code becomes GPL code, right? There is no takesies-backsies. There is no reversal." (2)* However, the issue is not just about the GPL as a legal requirement, but the norm that it sets. When a project moderator was asked if anyone had considered rethinking the project's open source status to control how it is used, he said that this would *"kill the project" (2)*, and that this would mean that the project gets less *"free help" (2)* and ideas. Another contributor stated that would require a lot of labor to do in a *"moral" (4)* way: *"rewrite the whole thing from scratch to make it closed source." (4)* Community members did not not seriously consider alternatives to open source licenses.

Participants also used the open source license as a reference point in reasoning about incorporating technical restrictions on problematic use. Leaders discussed an image recognition based content filter that would prohibit the software being used to create pornographic content, or embedding a visible or encoded watermark identifying the video as a Deepfake to enable people to distinguish between doctored and real footage. However, many participants believed there would be *"no point" (1)* putting in restrictions because the project's open source status means such safeguards could be easily removed: *"I cannot stop people [from] using my software for stuff which I don't agree with [... open source's] positive is also it's negative: [...] anyone can read all the source code and then can change any of the source code they want [...] whilst you can build stuff in to maybe stop your software being used in the way you want, someone [can] just rip it out again." (1)* Other participants believed *"forcing" (0)* such restrictions would require them to *"actively invade our user systems" (2)*, reflecting not only a practical but moral aversion.

"Forking" projects—copying the code into a new repository and working on it anew—is frequent in open source [282], which has the effect of distributing and decentralizing control [62]. This led another leader to believe that forking would lead to an additional, separate community without the ethical guidelines and content moderation they use: *"Let's say I built a load of limitations into my software [...] and anyone who used it, uh, would fall afoul of those filters. Well, what should happen is that the code would be forked and then everyone would start using the fork [...] And what effect does that have? It takes people to a version of the software, which doesn't have the ethical guidelines and doesn't have the moderation in place to make sure people aren't using for that. So you're kind of shooting yourself in the foot." (1)*

Another participant recalled when GitHub removed a project used by music pirates [61], leading to broad proliferation of that project's code (*i.e.*,, Streisand effect [133]) and expressed that restricting access would thus backfire. Another also believed this: *"If it was shut down, if the code would be deleted from GitHub, everyone would have it still on their computer and it would be easily find-able on the dark web." (7).*

Decentralized control in open source also makes some technical approaches to preventing harm more difficult, as one participant explained: *"Some of these server-based [deepfake] apps [...] actually have filters [for] nude pictures. [...] That's a different kind of setup because [...] they're taking photos that people are uploading then processing them on a server then spitting them back down to the user. So because of the centralized control [...] they could implement filters. I don't know that it could be practically implemented in an open source project that isn't server-based." (5)*

Finally, the transparency to examine source code provided for by the open source license was seen as an important resource for overcoming some types of harms. For example, a participant explained someone had embedded malware in a closed source app made using the original face swapping algorithm: *"he started putting a crypto miner in the program. [...] any closed source application like that in a relatively niche area has the potential for someone to put some sort of*

*illicit material in there" (2)*

**"This genie's out of the bottle": Technological Inevitability**

Many participants believed that because the original Deepfake algorithm is widely circulated, further development of Deepfake technology is inevitable, arguing that halting their own development work or other restrictions would only *"delay" (7)* development, but would ultimately be ineffective. One stated: *"if our project shut down today, deleted everything, there are other ways of [creating Deepfakes]. I mean, there are several other ways, uh, and you see them pop up, like I've [seen another app] and [another open source project], there's another piece of software and there are others." (4)* When discussing that their project had likely been used to attempt to influence an election, one project leader stated: *"if it weren't for [our project], they would have [another app...] It's not like the amount of work that it takes to make a face swap is far less than finding [our project] or one of its competitors" (2).* The same participant extended the alternatives idea from alternative projects to alternative individual contributors, referencing his involvement: *"In the end of the day we knew that that sort of thing was going to come about whether or not I participated in [the project]" (2).*

Some laws now criminalize non-consensual pornographic use of Deepfake technology [136]. Some participants viewed laws criminalizing the use of Deepfakes as naive given this inevitability, one saying those intent on unethical uses would not follow regulations anyway: *"Heavy-handed regulating is just going to hamstring us because there are countries and actors out there who just will do it [create Deepfake software] anyway, right? [...] If history has shown us anything, that if you ban something, it just goes underground" (6)* Another invoked a genie metaphor to argue for the irreversibly of technical progress and express distaste for regulatory action: *"I also don't believe in like, just banning something because it could be dangerous. It's just, first of all, it's not going to work. You know, this genie's out of the bottle." (10).*

Historian of technology Arnold Pacey framed the technological imperative which fuels this

feeling of inevitability as "the lure of always pushing toward the greatest feat of technical performance or complexity which is currently available" [202], and mathematician John von Neumann said that "technological possibilities are irresistible to man" [187]. Our participants appear to embrace this alluring inevitability, one participant referencing futurist Ray Kurzweil and then stating *"There's nothing that can be done to stop the steam engine that is progress. And technology, it's only getting better, faster" (3).*

Philosopher Daniel Chandler argues that surrendering to the the technological imperative "implies a suspension of ethical judgement or social control: individuals and society are seen as serving the requirements of a technological system which shapes their purposes", and that it is possible to abandon even "large, complex, interconnected and interdependent" technological systems, "given the political will" [53]. We see that our participants view their own role in developing Deepfake software as insignificant in the context of the wider progress of mutually interchangeable alternatives. They point to the proliferation of the original face swapping scripts before their specific project, and the broader idea of Deepfakes, as evidence. In a similar vein to the debate on nuclear proliferation [226], some participants framed these other parties as "competitors", and developing this technology as a race, thus making this needed widespread "political will" feel impossible. Implicitly, participants point out that to halt it all together, the political will to do so must be held by many uncoordinated open source, private company, and state actor developers of Deepfake software.

**"If I painted something offensive, you can't blame the paint manufacturer": Just a Tool?**

Some participants stated that they view the project as a tool, and that the ethics of any particular use case is solely up to the user, in line with views expressed by academic, public and private sector AI practitioners [201]. One contributor stated: *"You can't really blame the project cause it's like blaming the people that make the paint and the canvas [...] You can't blame them directly by no means." (4).* This participant then localized this sentiment to their project specifically by

comparing it to the image editing tool Photoshop: *"I mean we provide the tools, but then again, I mean, would you blame Photoshop if someone just put someone else's face on another's body? I mean, no! That's ridiculous." (4).* Others also employed the Photoshop comparison (which has also been discussed in past research [273]), stating while they believed Deepfaking has a greater ability to harm, the use of the technology is up to the conscience of the user: *"Face swapping is basically a more sophisticated application of, for example, using Photoshop to enhance the figure of a model. I think obviously it's more powerful and it has a greater potential to harm people, but I think the use of the technology has to be left to the individual conscience of the user" (5).* Others compared the project to to recent uses of long-criminalized psychedelics to treat depression, and cannabis to treat other medical issues, suggesting that it would be bad to *"hamstring a wonderful technology on the risk that a couple of bad actors will do something [bad]"' (6)*

One of the project's posted statements explicitly states that the project can be used for "good" or bad, a property it claims is common of any technology , which alongside views expressed above, reveals an *instrumentalist* view: while the way a technology is used may have moral implications, the "technology [itself] is neutral, subservient to our beliefs and desires; it does not significantly constrain much less determine them" [230]. However, as we will see in Section 2.3.3, some participants acknowledge that project's design can influence how it is used. Another participant agreed that changing the project's design could make certain uses less likely, even if not impossible, by implementing technological restrictions into the code: *"For people that [want to make problematic pornography] they're not very into [...] how it works. They just want the end result. [...] Right now you have to do quite a bit of manual stuff and you have to set up the whole environment..." (7)* Thus, he suggests that technological restrictions designed into the project *"could be a future idea that would stop a lot of people already" (7)* from using the software unethically, except for the *"very good programmers [who] will be able to take that [restriction] out" (7).* This participant reached a conclusion similar to many before

[126, 149, 150, 230] which we discuss in Section 2.4.2: *the design of tools make certain uses more or less likely, by requiring time and skilled labor to circumvent restrictions*. As we saw in Section 2.3.1, project leaders decided against restrictions for fear of their easy removal, but also worried that they may lead to splinter communities without the ethical norms we will now discuss.

**Setting and Enforcing Counter-Norms by Denying Support**

We saw in Section 2.3.1 that open source licenses shape views about developers' possibilities and responsibility for limiting downstream harm by presenting the right to use software for any purpose as paramount, but the project's leaders sought to set countervailing cultural norms to actively discourage uses they believe are unethical, without preventing such uses completely. There is a long history of open source communities setting norms outside those laid out by licensing, a process that Free Software anthropologist Chris Kelty describes as a "punt to culture" [140]: developers turning to persuasion, rather than strict, punitive control via legal or technical means.

The tactic of setting and enforcing counter norms is most clear in a public statement intentionally displayed as a *"very public policy" (2)*, which states that they intend their project exclusively for "ethical uses", and that it is not for creating "inappropriate" content. One developer for the project reflected that this is difficult to enforce: *"One of the points in our [public statement] is that [the project] is not for changing faces without consent or with the intent of hiding its use [...] Again, we can't force our users to do anything." (2)*. Enforcement, appears less important than articulating what does and does not count as harmful use in the eyes of the project. This has the effect of building consensus, which, in a distributed environment where projects can fork at any time, can be powerful. This tactic is also visible in the argument seen in the previous section that technological restrictions would make ethically undesirable uses *harder*, not as much a literal strict control as discouraging unethical use.

Leaders often expressed the view that denying valuable technical support [206, 263], to those

attempting to create Deepfake porn is their only way to *discourage* such uses, absent being able to outright *prohibit* them given their understanding of the legal dictates of open source discussed in Section 2.3.1 and 2.3.1. This is shown by the quote: *"So there's not a lot actively I can do. [...] But what I can do is discourage it and not [...] offer advice, and actively block people looking for that advice within forums and domains that I have control over." (1).* Project leaders recounted when they have banned people for soliciting help to create Deepfakes that contravene their rules, often after discussing the offending case privately amongst other leaders first. Another leader stated that refusing support is the "best" means of control they have:*"Best we can do is say, we refuse to support you" (2)*, going on to say *"if people are using it for that sort of thing, they're not going to tell us" (2).* Others framed this in terms of choices about their own labor, which fits squarely with open source notions of freedom: *"I don't need to teach anybody or learn how to put Scarlet Johansen's face on, you know, insert porn star here" (3).* Here, withholding of support became a matter of maintaining community, both in terms of who participates, what activities are acceptable, and how people choose to spend their time, which is not seen as in conflict with open source norms *per se*.

Combined, these efforts are having clear effects. Users of the software echoed the sentiment that the developers of the software are largely doing all they can to prevent misuse: *"I think there that they're probably doing all they can [...] it's not like they're going to be able to build like a detector or something for how the software is being used." (10).* The effect of these norms requires individual community members to take them as seriously as the users we spoke to did. Because these additional norms are not strict rules (anyone *can* use it for any purpose, per their GPL license), some weigh them against what they see as a higher purpose: the foundational norm of producing open source code. One project leader reported sometimes learning of pornographic uses of his software from crash logs, but reported overlooking this in favor of improving the software using these helpful logs: *"I try not to read what those are because they're not important for what I'm doing, but you could argue that I should ban people as soon as I see [them]. From*

*my point of view, I want to make the software better. So the crash report is useful for me. And as I said, I can't stop people using it for reasons I don't agree with, but I can discourage you." (1).*

### 2.3.2 Motivations for Ethical Action

Participants expressed intrinsic motivators for wanting to prevent harm, namely commitment to their own ethical lines and extrinsic reputational costs.

#### Ethical Lines: Consent, Family, Law, and Professional Standards

One leader described the creation of the public statement expressing the project-wide norms of acceptable use as arising from a kind of spontaneous agreement: *"We just all happened to be in the same place" (1).* However, participants explained how they arrived at this norm in a variety of ways: a commitment to consent and concern for the harm caused when it is violated, as well as a commitment to familial norms and professional and legal standards. Studies examining the motivation of open source developers on technical matters identify similar intrinsic or altruistic motivating commitments [13, 117].

Many participants demonstrated reverence towards the concept of consent. One participant spoke about how it is wrong to non consensually use someone's identity to sell products, saying *"you can't steal a celebrity's likeness to sell a product, right?" (3).* Another professional Deepfake creator created a Deepfake of a deceased person on the request of their relatives, but expressed ethical concern about whether this respects the deceased person's consent. One participant discussed how *"consensual pornography is completely up to the people involved" (2)* and a project leader echoed this:*"I don't have an issue with porn." (3)*, but then explained their own support for the blanket ban on asking for help creating porn because of practical and moral complications in ascertaining consent: *"It might be their wife and they have some weird [Deepfake] fetish. Okay. That's their thing. [but] It might be the neighbor's 12 year old girl that they got the hots for and have been videoing from a distance. No, [...] I'm not going to take the time to sit*

*down and [say] Oh, maybe there's a gray area." (3).* However, others believed that it is ethically permissible to create porn of someone else without their consent, because they believe sharing it is where most of the harm may lie: *"I think that's, it's okay to enjoy whatever you want, as long as you don't hurt other people with it, [...] obviously posting it online for other people to see and potentially for the person you don't have consent for, to find out that that will have a negative effect on them." (7)*

Others tied their personal sense of morality to how members of their family may react to certain uses. One participant initially said *"I really don't know how to define what's right and wrong" (3)* but then proposed a standard by asking *"would I show my mom?" (3).* Another participant stepped up a generation to suggest a litmus test to catch possible fake news: *"if you tell your grandma about it and you fooled her, and she thinks it's real, but it's a fake and it's saying something negative about someone else that's, that's not kosher" (11).*

Finally, others invoked professional and legal standards when discussing their personal sense of ethics. One participant who operated a Deepfake based marketing firm discussed a "very clear" line for his firm, informed by his experience as a photojournalist: *"We don't cross the line. [...] We follow things like [...] various journalism association standards and normal things you would follow if you're a Washington DC political correspondent" (11).* Another professional Deepfake creator declines pornographic Deepfake requests by explaining to prospective clients that such uses may be impermissible under law.

### Reputation

Past research shows reputation motivates open source contributors and influences their behavior: open source contributors actively promote their contributions to gain status [70], reputation is important one's contributions being accepted [123], and that job candidates and employers see contributions as indicators of technical skill [174]. Here, reputation motivated ethical action at the personal level for hobbyists to label Deepfakes as such and for professionals to attract

business; in the project we study to protect itself from censorship and differentiate itself from competing projects with perceived less ethical behavior; and the wider professional Deepfake community to escape the stigma of Deepfakes.

At the *personal* level, hobbyists strive for realism to show off that they are creating realistic Deepfakes, which calls attention to its fictional nature: *"if I could ever achieve [...] undetectable realism, then obviously I was gonna make a big [...] hoo-ha about it!" (6).* Another explained why most Deepfakes are labeled as such, reducing the risk of fooling people in his view: *"Truly cutting edge [Deepfakes] are presented in a context that highlights the fake rather than disguises it, which is no surprise as the poor sod who's worked on it would naturally want to draw focus on their effort." (8)* Similarly, professional Deepfake creators reported creating high quality fan-art Deepfake content to post online to demonstrate their skills, get exposure, and get business. These people advertise a Deepfake explicitly as such for reputational gain, and these participants believe this mitigates risks of fooling viewers.

At the *project* level, leaders have gone to great lengths to protect the reputation of their project, because it had been previously delisted from Google results, put behind a login wall on GitHub, and had members banned from their Discord because of associations with non-consensual pornography in the media. One leader reports that the project's public statements were in response to the project being delisted and blocked. He also worked with GitHub to remove porn and porn-related images from GitHub issue threads created before he led the project, and adopted a contributor Code of Conduct to defend the reputation of his project and as a condition for GitHub to remove the login wall from their repository. Another leader explains that *"We don't want [the project] to be identified as hostile [...] We want people to be able to find us and find the software without having to face a deluge of nonconsensual pornography" (2).* A user of the software echoes this, saying the public statements are a *"very good" (7)* idea because then *"the media doesn't think that there's a group of programmers just trying to create blackmail software. Then it might've been shut down by GitHub." (7).* We see that the leaders of the project

engage in activities to limit unethical uses of their software partly in response to enforcement actions by the platforms they depend on.

The leaders reported a feeling of unfairness, pointing out that another Deepfake project's Github page links directly to a porn website and its forums to provide technical support, yet it apparently has not faced the same restrictions or had to do the same work to maintain a clear reputation. At the same time, when one leader is asked how he'd feel if his project was used for something he disagreed with, he replied *"I don't think I'd feel particularly bad about it because I'm not naïve [but if something went viral with his project's name attached] that would bother me, because that would be an association with my product" (1).*

Finally, at the *professional community* level participants who were members of the professional Deepfake community expressed an interest in protecting the ethical legitimacy of Deepfaking as a practice. One participant who is part of a small community of highly-skilled professional Deepfakers said *"[it is] frustrating because everyone that I know that's [creating Deepfakes] is doing it for the creative possibilities, to explore the ethical uses of [Deepfakes]. And it's like, you know, it's an uphill battle because of the sensationalism, um, about Deepfakes" (10)*, further describing the competing open source project which promotes the creation of non-consensual porn as *"unprofessional" (.)* Another participant explained that this negative reputation is *"a large part why most of those within the community [...] tend to be rather hostile towards those who show up asking for tips on how to create [pornographic content]" (8).* One participant explained that they have attempted to rebrand: *"a lot of us 'Deepfake' artists have come around to preferring the term 'synthetic media' [...] leaving the stigma of "Deepfake" behind." (10).* A casual user of the software expressed empathy with professional content creators: *"It's an association no one wants, to have the effort put into creative works using the tech marred by the association with these less than respectable use-cases is certainly no fun for content creators" (8).*

### 2.3.3 The (in)Accessibility of Deepfake Realism

We found that Deepfake realism is prized, and some suggested that more people should have access to this artistic tool, while others argued that difficulty achieving realism mitigates societal issues.

#### Deepfakes for the "Everyman"

Participants celebrated that the ability to create Deepfakes is now broadly accessible to everyone, not just to those in academia or in companies with special training and technology. One participant stated: *"There's something quite thrilling about the everyman (sic) having access to the tools to create results that depending on hardware could be on par with what industry professionals might cook up" (8).* We note that the gendered term "everyman" betrays something participants did not address directly: that these are tools made and used largely by men. Some did, however, recognize the harms to women associated with misuse. We discuss this briefly in 2.2.2 and point to literature discussing intersections between gender and Deepfakes in 2.2.1.

Nevertheless, widespread access was seen as a self-evident good: *"Machine Learning is an incredibly complex process which generally is the remit of academics. And so my drive for developing [this project] is to basically take this kind of impenetrable area of computer science and try and make it as accessible as possible for people." (1)* Echoing this sentiment, a professional Deepfake creator speculates that the output possible from a competing open source Deepfake project is equal if not superior to the work that leading visual effects firms are capable of: *"I don't think there's another program that you can get open source that can do what [open source project] does. I imagine like maybe Disney and ILM [a visual effects company] have home-built tools that can compete with it, but I honestly don't think [they do]." (8)* This sentiment is crystallized in public statements on the project, which portray AI as exclusive knowledge, documented in arcane research venues, but that their project opened participation to all.

This impulse to "democratize" access to an inaccessible technology by wresting it from the

hands of an exclusive few for the benefit of common folk is an ethical ideal which sparked the Free Software movement [62, 237]. This is a different notion of democratization than those seen elsewhere: a minority of "Ethical AI" guidelines from companies and governments reference political ideals such as open dialogue, broad participation and wider principles of democracy [137], and private companies are increasingly co-opting similar political language when marketing their AI endeavors [47]. Interpreted in the context of this wider political landscape, some of our participants accept the possibility that their software is used unethically to prioritize an ethically charged commitment to democratization.

### Inoculation through Proliferation: More Deepfakes as Remedy

Some participants argued that the antidote to ethical issues stemming from Deepfakes, such as fake news videos or defamatory porn, is increasing skepticism and distrust of videos which will be brought on by the deliberate and increasing proliferation of Deepfakes into the popular consciousness, whereas keeping them *"locked away would do more harm than good." (8)* This sentiment is expressed by the leaders of the project, one saying: *"One good reason to promote the use of Deepfakes in satire and in various other areas is inoculation: teaching people not to just blindly believe what they see." (2)* By analogy to Photoshop, one participant explores a world in which Deepfakes are not widely known or accessible: *"Imagine a fictional world in which Photoshop as we know it today is something only accessible to a select few industry experts with a budget of hundreds of thousands if not millions. Due to the far reduced exposure that the everyman might have to the works that can be created with Photoshop they would be far less liable to question a doctored photo when seeing one." (8)*

In this way, participants argue that *"ubiquitous" (1)* proliferation of Deepfakes becomes the cure to the harms this proliferation may bring, by "inoculating" people: making them not trust videos they come across without further verification.

**Low Accessibility and Realism is a Safeguard**

The previous two discourses saw access to Deepfake realism a greater good or even a way to prevent harm, but disagreed: some argued that extreme ethical concern is unwarranted because the high effort needed to make realistic Deepfakes prevents some bad actors from using it for ill, and that unrealistic Deepfakes unlikely to fool people. For example, a minor contributor to the project speculated that: *"They're not making it more accessible, I think on purpose to weed out the people that don't know a lot about technology and just want to do it for bad intentions." (??)*

Another participant who Deepfaked President Biden with dubious realism stated that he thinks those with political agendas are unlikely to expend the effort required to make realistic Deepfakes: *"I put Biden as the Trololol guy [an internet meme] and you can look and it's not great, but it's funny, you know, and that's about, yeah, I don't think anybody with a political agenda of some form is going to put much more effort than I did into it. So you're going to be able to tell [it is] fake. So it's not like it's going to change the direction of a country or something like that." (4)* One of the project's leaders stated that though he wishes people would explicitly mark Deepfaked videos as such, he thinks they are implicitly marked because they are often low quality: *"I feel like it is clearly marked even if they don't put it in the tags, because Deepfake quality is not really there." (2)* The project's leaders *are* focused on improving the quality and realism of the results, however, so any ethical benefit of having Deepfakes marked by their low realism may not persist.

Most considered professional work to be quite distinct from home-made Deepfakes. A moderator of the project referenced the movie Avatar, lauded for its visual effects [197], to explain that convincing fakes have long been possible with a large production team, convincing home-made fakes will be rare. Professional Deepfake creators describe those dedicated to highest quality as a small community analogous to the early days of long exposure photography: *"you have to be a pristine technician in handling all the parameters to set up your camera and everything" (10).* Similarly, a user of the project said he's never seen a Deepfake that he thinks could fool

people, but high cost will prevent this for the foreseeable future: *"convincing higher-resolution models require exponentially more high speed video memory. As it stands this is not cheap at all, and won't get cheaper for some time still." (8)* Here, participants are assuming that technical prowess or access to expensive hardware aligns with ethical scruples: people who can overcome technical hurdles to create convincing Deepfakes are less likely to create ethically problematic Deepfakes.

While the sentiment *"I've never seen it to be done realistic enough to pose any sort of ethical issue" (9)* appeared widespread, one participant expressed fear about the project enabling widespread, indiscernible Deepfakes: *"If [this project] is that accessible and that, because computers will get better, everyone can do it on their phone and in a bunch of years. It'll be scary if video evidence would never be trustable anymore." (7)*

## 2.4 Discussion

### 2.4.1 Helpless to Challenge Freedom 0? Limits and Possibilities for Developer Agency

Many participants felt unable to control downstream uses of their software, given the dictates of Freedom 0 – a core principle of Free and Open Source Software which demands that users should be allowed to use the software for any purpose, and is a primary way open source "democratizes" [62]. Throughout the research, we saw that Freedom 0 was treated as an unquestioned default norm more so than an accidental effect resulting from a mere choice of license. Freedom 0 is so fundamental that it is even encoded into the platforms that projects depend on. For example, the code sharing site GitHub's license picker only points to licenses where Freedom 0 is protected justifying this with the pithy statement "An open source license protects contributors

and users. Businesses and savvy developers won't touch a project without this protection."[2] The strength of this norm meant that at times participants expressed either misconceptions about the implications of their license choice, or their ability to have chosen a different license, or mistrust that people would abide by alternative license terms. For example, the same leader who lamented that they cannot control what people use it for was involved in choosing the license which inscribes this relinquishing of control. This project chose a General Public license (GPL), a "Copyleft" license where publicly-available derivatives or subsequent versions of the software must be distributed under the same freely-released terms. The choice is indeed effectively irreversible without the consent of the many anonymous past contributors, but this leader did not articulate his own agency in the first act of choosing a license. Similarly, we also see Freedom 0 at work in the tendency to view technical controls in literalist terms, and therefore to find them ineffectual rather than norm-setting.

The norm of Freedom 0 underscores and elaborates other discourses like Tool Neutrality and Technological Inevitability, which also frame designers and developers as lacking agency. These discourses are also common in proprietary contexts, but there, the ability to choose among or create bespoke closed source licenses is more visible and common because there are other concerns (such as liability, financial obligations, or regulatory requirements) that make the need to limit uses (such as, to paying customers) more common and apparent. Where action is taking place in OSS, it is happening via other discourses, such as setting counter norms and making choices about where one's unpaid labor goes.

The project we studied was not prepared to question Freedom 0. However, Freedom 0 is situated in a changing field of claims and counterclaims about software ethics. This field has a long history, including the contentious term "open source" itself, which represents a change from the early days of critiquing of business practices that restricted access to source code [62], towards the promotion of "open source ideas on 'pragmatic, business-case grounds' " [251]. Just as prac-

[2]See: `https://choosealicense.com`; some licenses require the free sharing of resulting derivative code, which companies may desire to keep proprietary.

tices and licenses changed in this previous shift, it is possible that projects committed to Freedom 0 may be forced to respond to newer changes. For example, the Ethical Source movement, part of a broader reckoning inside and outside tech companies, was founded to participate in "giving [developers] the freedom and agency to ensure that [their] work is being used for social good and in service of human rights"[3]. This centers the *developer's* freedom to choose how the product of their labor is used, away from the *user's* freedom to use the software for "any purpose", with the goal of using licenses to foster that "make it easy for the user to do the right thing" [140]. This recentering does appear to call for a rethink of Freedom absolutism. Other developments have similarly recentered the importance of developer labor rights. The *Tech Won't Build It* movement "holds that workers developing AI/ML should have a say in how such technologies are deployed" [239], and the Tech Workers Coalition advocates (among other things) for workers to have a say in how the products of their labor serve "people, communities, and the environment rather than solely [...] profit" [4], aligning with the Ethical Source movement's framing of freedom. Whether commitments to Freedom 0 change in light of these broader changes, is a key question for the future.

### 2.4.2 Transparency and Accountability for Implementation vs Use Based Harms

AI systems can cause harm in multiple ways, and locating the causes of each harm on a continuum between *implementation* and *use* may be conceptually useful in debates on how to mitigate them. We define *Implementation harms* as those arising through code, algorithm or data problems that can be fixed without changing the intent, or use, of the software, for example through the use of "de-biasing" techniques to reduce bias in algorithms or training data [17, 56, 252]. On the other hand, we define *use based harms* as arising from a use which may *itself* be harmful, that no amount of technological fixes, implementation improvements, or more or better code will

[3]See: `https://ethicalsource.dev`

alleviate the ethical concerns with the software. Some make this point through satire, showing how dilligently followed "ethical" implementation fixes do not alleviate the patently harmful use of mulching elderly people [142]. Others find that corporate-backed AI values statements focus more on AI design decisions (implementation) than questioning the business *uses* which AI enables [108].

Our open source case demonstrates how harms originating from each end of the use-implementation continuum are differentially affected by the limitations of transparency [20]. Free and Open Source software offers accountability through individual traceability to specific lines of code. Grodzinsky *et al* wrote in 2003 that the "many hands" problem (*i.e.*,, collective responsibility [22, 84]) in software development can lead to "harm and risks for which no one is answerable and about which nothing is done" [195], but argued that open source enables individual-level accountability because "if a developer were to write irresponsible code, others contributing to the open source software would be unlikely to accept it. [...] Parts of code can be ascribed to various developers, and their peers hold them accountable for their contributions" [109]. This traceability indeed helps identify and rectify *implementation* harms that occur through code quality issues, as exemplified by our one participant's reference to a surreptitious cryptominer in an alternative closed-source Deepfake app, and the transparency that open source facilitates allows scrutiny which can help illuminate and mitigate unfairness in classification or prediction systems, arguably harder to accomplish when the model and data is proprietary [34, 221].

However, our findings show that Open Source has less power to support accountability for *use* based harms, because harm can be wrought not only from parts of code which may malfunction or be ethically inadequate in some way, but from the whole software package operating as its creators intend, but for a harmful use they did not intend. Notions of transparency in open source combine access for scrutiny purposes (referred to as Freedom 1 in the Free Software community [238]) with unconstrained use, circulation, and modification (codified in Freedoms 0, 2, and 3 [238]), a combination which allows *use*-based harms to proliferate. In our example

of Deepfakes, open source's transparency and unconstrained circulation can help such harms proliferate, allowing unscrupulous users to learn the relevant techniques and achieve their goals without the "friction" of rebuilding code from scratch. In short: open source's commitment to transparency of implementation allows strong accountability for implementation-based harms, whereas the same commitment to transparency allows use-based harms to proliferate, and absent a matching commitment to transparency of use which would make such harms visible, leaves it powerless to support similar accountability of use.

The risk of this openness aiding the proliferation of potientially harmful technology such as superhuman AI [41], and claims that open source contributors are unacceptably expected to abrogate control over the ethical impact of their creations [274] have been explored before, and we unpack how open source norms lead some contributors to accept similar risks. Others suggest that market logic will operate in open source development to prevent harm because "'good guy' AIs" will "out compete the malicious and incompetent" [119], echoing the trust that some AI practitioners place in market logic to diminish less trustworthy AI [201], but we instead find that this competition lead some participants to view ethically mitigating practices as futile (see Section 2.3.1).

Of course, implementation is not always cleanly divorced from use: the designers' intent, the affordances they implement, and the influence these affordances have on users change the likelihood of unintended use. For example, our participants disagreed whether the Deepfake software was a "Just a Tool" with harm determined exclusively by how its used, or whether technical restrictions on use (Section 2.3.1) or the difficulty of using the tool influence whether it will be used for harm (Section 2.3.3). Philosopher of technology Bruno Latour and others argue against the "myth of the Neutral Tool": that the design of technological artefacts (he uses guns as a more obvious example) encode "scripts" in their design which invite certain uses and behaviors while making others harder [126, 149]. To help unearth normative conflict in discussions on software ethics, we believe it is important to discuss harms resulting from a system's implementation, the

possibility for ethically questionable use, and affordances which allow the former to influence the latter.

### 2.4.3 Implications for "Ethical AI" Research: Assumptions of Downstream Control

Some companies and open source communities are wrestling with and increasingly accepting responsibility for downstream harms, as are some AI practitioners individually [201], but entrenched norms mean this is a slow and fraught process (see Section 2.4.1). However, mitigation strategies, for example Fairness Checklists, make reasonable assumptions about what the range of intended or possible uses are [3, 66], or weaker and often unspoken assumptions that software should not be shared, deployed or depended upon until algorithms are "sufficiently Fair". We term these **Downstream Control Assumptions**: that software producing entities *can control, know, or at least envision* how their software will be used through a mix of design intent, internal control over all the relevant features, postponing release of software, and contractual choices about appropriate customers.

For example, Google canceled its contract with the US Military to provide AI software which could be used to improve drone strike targeting (a use-based harm) after employee backlash [261] showing that Google can use contract law to exercise a fairly strict degree of control over how its proprietary software is used. This decision was politically fraught, but even before it was made, Google had a specific contractual relationship with a specific entity that it had the right to not renew, and was able to evaluate implementation harms (*i.e.*,, mislabeling images) by evaluating fitness for purpose with respect to that entity's intended use.

However, as our case illustrates, assumptions of downstream control and awareness are even weaker, in both a legal and normative sense, in open source. Freedom 0 licenses *legally* dictate that contributors *may not* exercise control over how it is used, thereby enforcing the broader *norm* (see Section 2.4.1) that they *can and should not* be held responsible for downstream use-based

harms. Open source software often has diffuse or often unknown users, and code is often freely remixed into other products [282].

Since these assumptions are so entrenched, our case suggests that "Ethical AI" research and design interventions would benefit from being explicit when making and finding ways to work effectively under loosened Assumptions of Downstream Control. "Supply chains" (the series of steps by which raw materials are converted into and delivered as a consumer product) are a construct which may help locate ethical decision making within business and community relations, and explore how different supply chain arrangements yield different outcomes. Supply chains can help reason about upstream [220] and downstream harms in [83] in offline contexts, and the UN has published actions companies should take to mitigate human rights violations in supply chains [222]. The supply chain concept has also been transferred to software [12, 199], and software ethicists have theorized about responsibility for downstream uses of software, for example arguing that "If proper precautions are taken to limit the distribution of [hacking software], the downstream uses are constrained" [275].

This raises similar questions in other ethnographic contexts. Guides for "Responsible" use of general-purpose AI libraries often assume use(r)s can be known beforehand: guides for the general-purpose and widely-used open source ML framework TensorFlow ask "Who am I building this for? How are they going to use it?"as a crucial first step for considering ethics when using it to build other things [78]. Are TensorFlow's ethics options different or similar to the smaller use-specific project we study? In the private sector, how do far upstream actors, like ML-as-a-service companies or ML-enabling GPU manufacturers, see their responsibility and the choices available to them? Whether researchers are studying open sourced technologies or not, making explicit whether possible uses are known or unknown, and where in the supply chain possible harms or mitigations are proposed, and the limitations this may bring, can expand and strengthen AI ethics scholarship by surfacing new points of connection and action along that chain, and opportunities for ethical action under these limitations.

## 2.5 Conclusion

In this chapter, we have examined how a community a deepfake tool understands their responsibility and agency to prevent downstream harmful uses. In addition to beliefs about technological neutrality and inevitability, we find that notions of "Freedom 0" encoded in licenses also set broader norms serving to disavow responsibility for downstream harmful use. We propose a continuum between harms from implementation–like bias in models– and harms from use–like the creation of deepfake porn, and argue that open source development contexts allow greater scrutiny for yet little visibility into the latter. We discuss how assumptions of downstream control are often implicit in "Ethical AI" discourse, but outline alternatives for cases, such as open source, where these assumptions cannot be safely made.

# Chapter 3

# Dislocated "AI Supply Chains" and Ethical Disavowal

Work in this chapter was originally peer-reviewed and published in the SAGE journal *Big Data and Society* in 2023, with coauthor Dawn Nafus [267].

## 3.1 Introduction and Background

Many big technology companies are building responsible AI programs[1] [137], but those "owning" these programs are limited in their ability to create change, resulting in varying levels of efficacy [179]. Even those without designated ethics roles are called to follow responsible AI guidelines [137], checklists [165], and other processes [233]. Outside of the biggest companies that build and deploy their own user-facing systems, many engineers operate at arm's length from their firm's immediate customer, who might themselves be multiple steps from a live deployment. How is responsibility and agency socially organized for AI practitioners in these distributed ar-

[1] "Responsible AI" as opposed to "ethical AI" appears to be the more common term. Our own use of "responsible AI" denotes our commitment to feminist theories of technology [115], where ethics cannot be removed from the question of "to whom?" does one owe a response. We sometime use "ethical AI" where context makes it appropriate.

rangements? What can be done in situations where responsibility is framed as checklist work, and where this work risks falling through the cracks between actors?

We investigate how AI practitioners scope their agency and responsibility to address possible AI harms. Our participants described situations where they were asked to account for harms their systems may enable, yet saw those harms as beyond their agency, capability, or responsibility to address. We were struck by the deeply dislocated sense of accountability, where acknowledgement of harms was consistent, but nevertheless another person's job to address, always elsewhere. We suggest that the software engineering ideal of modularity, and the divisions of labor it enables, re-inscribe a belief in software production as supply chain, where developers recognize their dependence on others' code much like a shipment of goods: as necessary supplies, but not where a deep collaborative relationship might develop. When harms were recognized, it was usually through social locations cross-cutting or separate from the "supply chain." We argue that these same cross-cutting locations can be used to rebuild responsible AI practice to recognize the limitations developers feel, while building inter-organizational linkages that enable societal and commercial value.

Other work has shown that engineers do not see business relations within their scope to consider [201]. [108] showed that many responsible AI programs scrutinize AI system design instead of questioning the business purposes these systems enable. Familiar responsible AI interventions, like checklists, model cards, or data sheets ask practitioners to map their technology to its end use, attempting to bring "out of scope" harms back in scope. We show how existing realities of software production work against this, catching developers between countervailing cultural forces.

The software engineering notion of "modularity" refers to a specific technical practice and the broader, inseparable cultural beliefs, epistemologies, and organizational arrangements it mediates and reinforces. Tech firms use metaphors of modular, containerized work to describe both code and teams of coders [113]. Modularity has been a staple of software development since

the 1970s, where large software systems are decomposed into smaller, self-contained parts, so one can control parts of a system without needing to address the myriad details of the other parts [231]. This "information hiding" [204] buries "the complexity of each part behind an abstraction" [27, p. 64]. This facilitates a division of labor and the matching of individual skills to specific tasks [231] by separating concerns of different workers [77]. In practice, modular software may need fewer repairs, and may be easier to repair, but software can also be *too* modular, perhaps due to error-prone and calcified inter-module interfaces [141]. Nonetheless, open source projects strive for modularity to make their codebase understandable [162], and professional software engineers see improved modularity as a benefit of refactoring their code [144].

This divided labor, inscribed in code itself, has enormous cultural and social implications. Modularity's apparent simplification facilitates the presence of "many hands" who are harder to keep accountable [194]. The problem is more than many hands, however. Modularity sets the stage for a refusal to accept a relationship between "us" developers and "them" technology users, let alone other affected citizens [178, 247]. Others have noted that modularity is an epistemic culture (*i.e.*, [52]) that cultivates a capacity to "bracket off" [168], even when human beings are bracketed off, not pieces of code. This makes it an everyday form of the modernist fallacy of the separability of society from technology [148], separating code from harms it enables. It is an example of the social organization of ignorance [214], where the focus on one thing (the workings of a single portion of code) yields ignorance of another (the activities of other developers and users). This ignorance is not total, but situational: our participants were aware of harm, usually when outside of their role as a software engineer.

While other factors, including crude profit incentive, deepen this dislocated accountability, modularity is a touchstone of technical practice that serves as a lens through which these other matters are framed. Developers imagine their work as an extended series of modules that form a chain, as if the whole were a summation of parts. They also imagine that any particular piece of code is embedded in other code that is "near" or "far" to the general public (see Figure 3.1).

By extension, entire organizations are also seen as "near" or "far" to end use, because organizations package up code to be "released" to other organizations. These are metaphors drawn from logistics. More than a metaphor, they also constitute the relations of logistics, from the obfuscation of distant labor practices to the security concerns that arise by not looking inside the "container" Hockenberry [121].

Here we focus on how the metaphor also defines other relations (business, personal reputations, user experience, etc.) as not part of the chain, but as a kind of secondary background. These "secondary" relations nevertheless hold things together in a different way. [50], for example, follows Latour's 1999 "chains of translation," to examine data chains that tie the precision agriculture industry together in recursive and contested ways. While developers imagine supply chains as a series of upstream and downstream modules, like so many cargo containers awaiting shipment, Carolan's work suggests that chains can also work differently, where the links are not as discrete. Sociotechnical relations might occupy multiple social locations and cultural logics at the same time.

The links in a chain form a boundary of some kind, making responsibility "a boundary-crossing activity, taking place through the deliberate creation of situations that allow for the meeting of different partial knowledges" [247, p. 94]. We argue that asking developers to anticipate every conceivable outcome by diligently following elaborate checklists as if they occupied a view from nowhere (what [96] call "metadata maximalism") does not portend a meeting of partial knowledges. We take a located accountability approach that sees "systems development as entry into the networks of working relations" [247, p. 92]. In this context, that means asking developers to soften the view that once it is out of "my module" – the place that appears to make total knowledge possible– it is out of their control. Instead of metadata maximalism, we argue it is more effective to find and acknowledge where working relations can or do exist, and where no single party has total knowledge or control. This is where developers can bring their partial, situated knowledge to bear. Even if technology use cannot be fully anticipated or controlled [147],

crossing boundaries between "modules" can and does reduce ethical debt [i.e. 89]. In this work, we identify key social locations that could create better points of boundary-crossing to reduce ethical debt. We conclude by suggesting that if accountability depends on the ability to critically analyze one's own social location, so that a developer has a better sense of to whom they are accountable, and what it is they owe others in different parts of the chain, a thorny question arises: what kind of critical reflection, engagement, or questioning of modularity can be expected given that modularity is itself a dominant form of social relations, and being located within it involves an injunction to reject the very notion of located accountability in the first place? We suggest three potential paths forward, depending on how deeply one is prepared to question modularity.

To conduct this study, we recruited using public emails and existing contacts, alongside paid services and snowball sampling to seek views from those working at various points in the AI supply chain, across different modalities of machine learning (i.e. computer vision, language processing, etc.) and application areas (i.e. military, manufacturing, medicine, etc.). Our participants were not directly in the same supply chain such that we could trace a single component through it, but they did reflect patterns in what it meant to be "upstream" and "downstream." Our 27 participants were primarily in North America (16), and Europe (9), with one each in Asia and Africa. Private sector participants worked in eight companies ranging from startups to established smaller companies to large multinationals. Four researchers from three universities participated. Seven participants contributed to six open source AI projects, sometimes as part of their employment, sometimes outside of it. Many had ML-related graduate degrees; job titles included: Machine Learning Engineers, Research Scientists, Developer Experience Researchers, System Integrators, and Project Managers. All identified as men except one woman, reflecting disparities in the AI workforce. Each were invited to a semi-structured recorded teleconference interview, which were then professionally transcribed, except for one participant who preferred that we take notes. Most interviews lasted an hour, but were as short as 30 minutes or as long as two hours. After asking about their background, daily work tasks and projects, we asked how

they thought the system they are working on may be used or misused, where they saw possible harm, and if there was anything they wanted to, could do, or currently do, to prevent it. A variety of interstitial documents accompanied our analysis. The first author wrote a descriptive memo after each interview including observations on how the participant described their agency on ethical questions, and added to a running analytic memo documenting connections between participant accounts, and categorized quotes representing these connections iteratively. We also produced a table to reassemble the emic "supply chain" metaphor, which allowed us to identify patterns in how each participant positioned their work on a spectrum from "general purpose" to "specific use." This became a resource for examining how the chain inflects views on responsibility. Our different positionalities helped us think critically about modularity, both from the standpoint of someone within computer science trained to see it as a valuable technical and social practice (omitted), and as someone trained to first see its epistemological shortcomings (omitted).

In the next section, we illustrate how a distributed AI supply chain limits developers' sense of agency and responsibility. We then show the various ways the supply chain is reproduced in practice, alongside the social locations outside the chain that create space for responsible action to be taken. We show how the confluence of the two shapes the ethics work that is and is not done. Finally, we present three potential interventions, depending on one's view about whether modularity is an ideal to be preserved or a problem to be overcome.

## 3.2 Views from Up and Down the AI Supply Chain

Outside of the largest technology companies, complex inter-organizational relationships are at the heart of building AI [250]. For example, computer vision used in a power plant's surveillance system to detect a person at its perimeter might begin life published as academic research, further developed and made freely accessible in an open source library as a pretrained model, later requiring *in situ* training when deployed to work with the plant's existing hardware and software

by a systems integrator. It might be further adapted if the plant has the requisite expertise. Thomas [250] observes that by 2018, computer vision professionals expected to not need to build systems from scratch, with open source tooling and pretrained algorithms available to "kick start their work," and find a role somewhere in the chain. The persistence of a chain metaphor is notable given that software development professionals have shifted from linear "waterfall" production methods to nonlinear, iterative "agile" practices [112, 121]. Chain metaphors come back into play precisely when developers imagine their scope of control, which they believe is limited by when a product is "released" by one organization and used by another. They also believe that control over their system's impacts increases as possible uses of the released system narrows, as it is adapted to fit a particular end use.

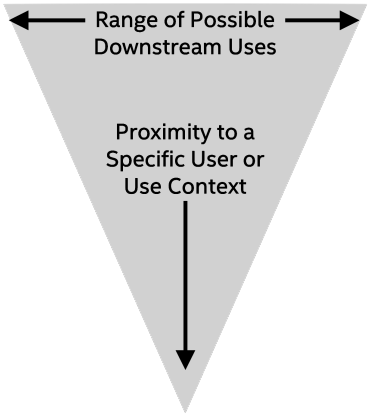



Figure 3.1: Work closer to a specific end-use context is perceived to imply a narrower range of possible (mis)uses.

Higher in the AI supply chain are supposedly general-purpose research outputs or tools, such as an academic ML researcher relaying his enjoyment in *"discovering generalized infrastructure components that are missing from people's workflows" (19)*, where *"the application domain you pick can be potentially endless." (19)* This endlessness gives this person a sense of value and prestige, while the ability to control impacts does not. Separability between the optimization procedure and what is optimized sustains the belief that optimization tools are "general pur-

pose", creating "endless" possibilities. From the top of the supply chain, the generality outshines the fact that there is a purpose of some kind, and that purpose precipitates some outcomes over others. Asked if there are ways his project could be used that would concern him, he answered: *"nothing that would concern me [except] general ways in which you can abuse machine learning. [...] I don't think it does anything that can be abused relative to what you could do normally with any machine learning algorithm." (19)* He extends the separation of the optimizing code from the optimized code to the people involved. This person at no point mentioned a "who" that might use his tool, suggesting he does not imagine there to be a social relation of some kind. He only imagines other inert containers of software, enabling him to normalize harmful machine learning practices as a general matter of course, or theoretical possibility, and not question his participation in it or his choices about who he allows to access his technology. His direct contribution to the "optimization" of harm by enabling it to occur in a more technically optimal manner is thus invisible. One might call it an uncritical technical practice [9], where incuriosity about the other persons' "container" in turn leads to an incuriosity about why he is spending his time optimizing "the general ways you can abuse machine learning." Indeed, people working high in the supply chain were particularly prone to employ discourses of technological neutrality [i.e. 272], referring to what they make as even more general purpose than the proverbial dual use gun: *"I make a piece of equipment that makes pipe, somebody bought my pipe making equipment, and made the barrel of guns. I don't know how I stop [harm], because I didn't make the gun." (8)*

This view is also situated in a neoliberal economic context where *not* having relations or obligations is a dominant model of appropriate economic behavior [49, 105]. Unlike gift economies or other economic forms that constitute staples of economic anthropology [209], the dominant narrative of economic exchange here is that there are no social ties after the exchange takes place. The parties are *quits*, with no further obligations to one another. This stands in stark contrast to the competing notions of responsible AI development found in the indigenous data sovereignty movement [51], where care and the building of relations is central.

In the middle of the supply chain lie partial systems like performance benchmarks or pre-trained models, designed to show off accuracy, speed, or ease of use, as "kick starters" [250] for others' future finished deployments. These contexts make upstream dependencies and downstream responsibilities more visible. For example, another engineer used *"a composition of already existing components" (16)* from an open source framework and models to develop machine translation *"benchmarks," (16)"showcase[s]," (16)* and *"demo[s]," (16)* which he also made available as open source. Because he did not build the framework, he stated *"it's a part of open source project so [...] we are not taking the full responsibility for the framework itself," (16)* downplaying whether he had any choice whether to vet it for problems. Looking downstream he stated: *"there is a very little interest in the actual... meaning of translation, but rather [more interest in] the performance numbers," (16)* like translation speed or accuracy. Because the output is not considered a final matter with real consequences, he does not consider it his job to address biases: *"I don't believe that anyone will try to prove that, hey, the output is biased." (16)* While he was somewhat concerned that his company's logo would be attached, he expected the next person in the chain to know to address it, which re-rendered it as a "general" problem: *"there is always a risk that the translation can be biased." (16)* He points to the least "general" actor in the chain as the site of responsibility: *"I believe that the final responsibility lies at the client's side who is finally deploying the actual service." (16)* He frequently used passive voice to describe decisions that he could have made otherwise, for example, *"the data **was taken** from official available sources," (16)* and *"existing components, **which are** packed and prepared." (16)* These felt like statements of fact, not attempts to be exculpatory. The participant began the interview apologetically, explaining that his *"very simple" (16)* project provided little for research on ethics.

Lower in the AI supply chain, an AI model is integrated into "live" software. Here, harms are closer and more visible, but managers still considered it a virtue for software engineers to be able to focus on their technical work, without interacting with those using their software. For

example, a tech lead at a company building VR services for defense industry clients explained that ethics are *"a concern to me because there could be flaws in the code, security risks, quality risks, and effectively, if anything goes wrong, it looks bad on us." (7)* Nevertheless, he talks about the separation of engineers from colleagues that handle customer interactions with relief that he *"kind of get[s] to turn a blind eye to certain social aspects" (7)* because *"we have program managers that tend to be the buffer." (7)* He says sometimes he gets pulled into customer conversations but they are improving the process to make sure *"I'm not involved, because frankly, I shouldn't be." (7)* If software engineers building the software might have an issue with their work being used to train military drone pilots, this separation insulates them from intimate knowledge of this use.

Downstream in the supply chain, the design affordances that limit use are more acknowledged. This participant was confident that his *"app isn't so open ended that it can just be used [...] by accident in a different way," (7)* noting it would take some reverse engineering to use it nefariously. But he is uncomfortable with his upstream dependencies, Facebook's Oculus: *"we're kind of putting our foundation on sand" (7)* because *"the platform [...] is owned by Facebook [which[] recently had a pretty bad day [participant referencing then-recent congressional testimony]. So frankly, we don't trust them." (7)* This raises the real possibility that he may be vulnerable to having to pay his supplier's ethical debt [89]. Looking downstream, he is also aware of the care that needs to be taken with respect to which customers he does business with. He states, *"There's always going to be some level of let's say customer qualification," (7).* Discussion of customer qualification did not occur higher in the supply chain.

When people talk in terms of "getting to turn a blind eye" to consequences, and normalize harmful actions as a pervasive yet unconcerning matter, we have a form of social organization that creates a partial ignorance of customers and suppliers. To extend the logistics metaphor, these developers imagine themselves inside the container, not piloting the cargo ship or even developing the software that coordinates supply chain systems. [211] points out that in supply

chains of physical goods, companies still struggle to gain full visibility into their networks of suppliers and labor or environmental conditions in part because the software that is supposed to create that transparency is as containerized as the goods and services it is meant to monitor. In this sense, the use of the supply chain metaphor is no coincidence; supply chains are a sociotechnical system of partial, selective sight [211]. This "view from nowhere" [115], then, is not a god's eye view, but a view from within a digital cargo container that knows little about where it heads. It is both difficult to know, because of the many hands problem, and there is little desire to know, because of the social organization of modularity. As [241] reminds us, claims that technologies need to be set in some context already tell us about the context they are, in fact, in: one believed to lack social relations. Here, modularity creates the numerous ways that responsibility is not to be found "here" regardless of where "here" is. Context is perennially displaced to elsewhere.

## 3.3 Crosscurrents Within and Against the Supply Chain

In this section, we discuss ways that the supply chain is reproduced, and the ways that people have to step out of the chain to prevent harm, whether in institutionally sanctioned or unsanctioned ways.

### 3.3.1 Reproducing the Supply Chain

Divisions of labor, an important purpose of modularity, create the cracks through which responsible AI actions fall. It is remarkable that relationships themselves — acknowledging the effects that one person has on another — is seen by our participants as an act of labor that can be divided between people and handed off. This is neither a natural nor obviously normal state of affairs, as in other contexts the very notion of it would be utterly rejected [see 158]. In this context, however, to divide labor is so naturalized that participants expected relationships to either be rendered into a task, or to not exist at all. One participant explained that no one tasked him with

doing ethics work, so he doesn't do it: *"I don't have time allocated during my normal week to think about [...] responsible AI. This is not part of the work, at least not the part that someone would tell me from the top to worry about." (4)* There was often consternation about who would do an ethics assessment. A user experience researcher stated that ethics assessments are often filled out by software engineers, and that *"it was not my role" (2)* to do it. This posed a problem to him, because there *"might be value in somebody who talks to customers i.e. me, filling it out versus an engineer." (2)*, echoing work showing that separating concerns between UX and AI work is difficult [246].

Status inflects divisions of labor. To the extent ethics was recognized enough to become a task, it was a task often seen as mere details. One participant filled out a privacy questionnaire for his team to use an existing dataset to build a speech recognition benchmark, and felt the questionnaire asked for a lot of seemingly immaterial details his team was unconcerned with: *"It wasn't that easy to get through all the sections [of the assessment...] there were some questions about how the storage is secured [...] a team member of a research team or engineering team is not aware of [that] – it depends on IT support and configuration." (12)* Others simply handed the work off to contractors or junior employees, as a form of administrative labor no one else wanted to do. This is hardly a meeting of partial knowledges that would be suggested by taking located accountability seriously. Instead, it follows broader patterns of status between work on the model versus data [225], and in programming generally [62]. Another university-based participant emphasized that he was encouraged to focus on results, which did not include the resulting societal impact of any kind :*"It's not like when we're presenting [our research at a conference] they ask you [...] what ethical steps did you take [...] Usually they just want to see your result." (5)* These divisions made the authority to decide questions of ethics ambiguous. One participant building body scanning technology explained: *"several questions [on the ethics assessment] are focused specifically on a machine learning AI statistical model, where many of the other questions are more around the broader product and business. So that was confusing,"*

*(9)* because making those assertions felt like an overstep of his own authority.

In addition to the division of labor, the pressure to "scale" to ever more data, users, and customers deepens the sense that others in the chain are unknowable and unconnected. For example, a VR service tech lead was concerned that while most of his current customers have been *"physically met by one of our team at this point, that doesn't scale" (7)* as they build a service company. Another participant discussed a deep collaboration with a customer to build an AI system on the customer's site, but felt unable to know what the customer later did with that system, as follow up work was believed to not scale, because it required labor to do it. Similarly, another participant said: *"So right now, I know the clients. And we don't have clients [who do harmful things]. But in the future, once we go public you won't be even able to control that [... with] 10,000 clients – I don't know how many clients we'll get [...] It can be difficult to track [...] what they do with the system." (4)* His careful knowledge and consideration of his clients, the metaphorical glue between "modules" of the supply chain, is the very thing he also would have dismantled in his (and his company's) ideal future of broad adoption.

While these participants saw scale as a desirable state that creates a regrettable limitation on attention, others thought it legitimized not doing ethics work at all: *"our company is so focused on growing and scaling with users that ethical AI is not really [...] a big concern at this point." (6)* Others thought this would create friction and lose customers: *"If you bring [ethical AI] for every other use case and every other customer, there is already a lot of customers that we are losing [...] I don't want this to create a bottleneck for our customers" (11)*, and even a limitation on technological progress itself: *"there is going to be hundreds of thousands of industrial uses of AI [...] But if we start limiting ourselves from doing so because of ethical concern then it stops progress of so many developments." (11)*

No one in our study articulated of a specific reason why one would want to scale; it was as if this was axiomatic enough to go unsaid. As [113] have argued, "scale thinking" is linked to modularity and capitalist impulses, and is also its own perceived moral imperative that cannot be

explained by economic or technical practices alone. These research participants are articulating the precise, embodied moments when scale becomes indifference: moments where conversation is severed, where the investment in care relationships wanes, and when context is no longer something one is a part of, moving from a situational awareness of harm (see also [164]) to a distant matter that needs "tracking." Participants invoked "scale" as a way of describing the removal of personal relations, as if it were impossible to know the motivations and desires of one's customers beyond individual personal connection, forgetting that there are entire business apparatuses designed to do so, like market research, customer management, or corporate auditing. What participants are expressing here is not a straightforward practical fact, but the way that notions of scale create a remoteness from reality that makes it possible to not see harm [107]. Notions of scale render "technical systems as commodities that can be stabilized and cut loose from the sites of their production long enough to be exported *en masse* to the sites of their use" [247, p. 95]. They reinforce the distinction between inside and outside a company, and create an important site of cutting a technology loose from its creators.

### 3.3.2 Acting Outside the Supply Chain

Social ties are not nearly as severed as the dominant discourse suggests. Participants were located in cultural logics that produce connections and responsible actions outside of the imagined triangle in Figure 1. Some of these activities are also the glue that holds the economic chains between organizations together, yet developers still saw themselves as stepping outside their supply chain role to act responsibly.

For example, being "customer-centric" was an explicit corporate value in many participants' workplaces that required them to understand how customers interact with their software to increase product satisfaction. User experience design plays a key role here. One participant led his team in a brainstorming session for their product to allow users to scan and monitor their body composition over time, which he felt was enabled by a shared and authentic *"passion for the user,*

*for the customer." (9)* To this end, they made design modifications in response to feedback from pilot studies with users, framing this as putting the customer's needs first: *"We recognize [health and body composition as] a very sensitive thing [... we're ...] focused on solving problems for the customer." (9)* While paying customers are often the privileged "humans" in "human"-centered design to the exclusion of other affected parties [205], specific anticipated users creates a connection point between commercial incentives and better or worse societal impacts, even if these were proxies for relations rather than direct relations themselves.

User-centered design connects designer and engineer to (imagined or real) user, but mechanisms like licensing connect customer and supplier, especially further upstream. One participant's company released its machine learning framework both as freely available open source and as a download available only after signing up with an email address. Of these very different relationships, the participant preferred the second method because *"we can be far more in touch with our customers. We know who they are, we can email them, we can make that more of a community." (8)* Being "in touch" clearly has economic value that notions of scale deny, but also holds potential to surface awareness of things that can go wrong downstream.

Marketing is another exchange point between actors. "Ethical AI" was seen as a marketing advantage, with one participant suggesting that it is a *"very, very good influencing tool [where] users might choose [our company] over the competition." (1)* Another believed that responsible AI can be used to win sales: *"the first thing that comes to mind is [...] how to earn as much as possible, right? [...] this Ethical and Responsible AI, [we are in a] world that using these terms could only help you, right?" (4)* Whether fortuitous alignment or crass co-opting, participants believed responsible AI efforts serve as a market differentiator, where companies can win business by helping their customers avoid ethical debt and the reputational costs it potentiates.

Similarly, engineers stepped out of the modules they build when thinking about how companies' ethical mishaps affect their own and their company's public reputation and profit. One participant relayed that his company had canceled a contract with a customer company which

was using his team's software framework in a widely-reported unethical way, and suggested why this happened: a *"public perception of your moral compass [...] has a direct impact on your bottom line [which...] makes company owners stand up and do something different," (2)* namely, sever relations downstream. Participants directly associated with potentially harmful projects also feared personal reputational costs: *"Some things can have uses that you don't intend, and that you don't want [...] to come back to you." (13)* Concern about reputation seems the most direct acknowledgement of the impossibility of fully disconnected, modularized work. Developers know the impact of their creations will follow them or their companies when others believe it was their job to control the problem, even when they do not.

Reputation and customer value are not new frameworks for legitimizing ethics work [179]. We should not interpret concern for reputation or attention to market value as always an indication of empty veneer. "Reputation" is the language through which social relations are acknowledged in a context that has an exceedingly thin vocabulary for them. Interviewees did not veer too far from their professional personas, where flat affect is the norm, and private beliefs are expected to be contained into their own separate module. While we have little evidence, we suspect that for some, concern for reputation might reflect deeper notions of obligation for which there is no local vocabulary, while for others it might solely reflect concerns for economic consequences, while for others still, the two concepts might not be separate at all, and economic penalty might be taken as a sign of social disapproval. When participants wrestled with the problem of conflicted interests, the motives for repuational concern were questioned only when it came in the guise of other people. For example, one participant says he hears the term "ethical AI" from *"C-suite kinds of people," (2)* but questioned whether this was a *"buzzword" (2)* or whether something was *"actually happening." (2)* While he believed his company doesn't want to *"be a party to any inhumane usages of AI technologies" (2)* by downstream customers, he said they also want to *"make money. And sometimes those are cross purposes." ()* Similarly, another participant framed Google's treatment of Timnit Gebru as something that *"communicates that they*

*care about ethics [only] to a certain point." (6)*

There were also instances where corporate rationales were not what motivated ethical action. The participant working on the body scanning project, for instance, emphasized that his team's positive group dynamics was what made it possible to talk about ethics concerns by studying each other as pilot users, having their own bodies scanned, and sharing their intensely personal reactions. For this participant, ethics discussions were an exercise in vulnerability, and responsible design meant a powerful obligation of duty to one's colleagues and friends in the position of "user." While the technique has its limits [32], it is arguably more potent than hollow onstage rhetoric [i.e. 100, 228] of "passion for the customer" or "human-centered design."

Ethics issues are not so easily disavowed when asked about work by friends and family: *"It sometimes gets hard when other people ask me. [...] 'What do you do?' [...] 'Oh, I kind of - I work in the AI workspace?' 'Oh, so you're getting people killed and assassinated through - with drones [...]' and it's like well, how much am I involved in that? [...] You can't say it's not true because it is true. [AI] is used for that." (2)* Work on a "general purpose" framework did not allow him to unsee harms when called to account in social contexts. Others talked about wanting more from their employer. One person noted that they could not necessarily say whether their framework was being used by the US Army, and this not knowing was itself a kind of harm: *"that's one thing I would really like to be informed, when my software is used. Where? For what purpose?" (4)*

We also heard of developers exerting a soft form of agency and resistance when their moral compass made them uncomfortable with assigned work [276]. One participant's company's client asked them to track the actions of garment workers. Having inspected the training data the client provided, she stated, *"It was a little sad looking at videos. They work from 6:00AM in the morning to 9:00PM at night." (3)* She said that even though the client called the project "object tracking," she was concerned that it would amount to algorithmic management: *"the algorithm that we're using is basically looking at people's motions to figure out what exactly*

*they are doing. So, sometimes [...] they're just taking a break. You're just telling the system that this person's not doing anything." (3)* She described how her team deprioritized the project until the client pulled away: *"[it was] not a project that any of us really wanted to work on. Thankfully it didn't go anywhere." (3)* This is softly subversive [276], in that subversion was undertaken through inaction rather than overt action. It is remarkable that otherwise elite and well-resourced AI developers nonetheless still feel they must resort to weapons of the weak [i.e., 227]. Whether caring for relations among cowokers or friends or for workers on a video who appear to be exploited, there is a quality of off-stage norm-making that is not encapsulated in official talk of "customer orientation" and responsible AI transparency interventions.

In practice, these crosscutting impulses to divide and connect lead to particular ways of handling responsibility and particular areas of priority. What does get attended to are matters of widespread public concern that can be encapsulated into a module of work without introducing friction into the development process. High-profile ethical lapses like racial and gender disparities in computer vision [46] and marquee regulatory action such as the European General Data Protection Regulation provide a shared social location, from outside the supply chain, from which to recognize some harms, but not others, within it. Bias might be measured statistically, but not questioned in other ways. For example, one participant doing AI research for the military was concerned about the mathematically-identifiable biases within the weaponry, saying, *"I think the whole issue of bias and its societal and ethical implications is terribly interesting and we don't have as much conversation, particularly with cyber weapons, as we should." (14)* Measurement fit the module, while any bias in the choices his customers might make about who to point weapons at did not.

This social configuration leaves us with an odd bimodality. On the one hand, prominent dramas about social harms embroil the careers of executives in congressional hearings, while on the other, contractors are asked to do "the paperwork." In a hollow middle, some limited actions do take place. Disparities in accuracy rates are often checked. Offstage action, like slowing work

or meaningfully caring for a colleague playing the role of user, remains invisible, like a shadow responsible AI workforce with little connection to checklists, transparency, or customer vetting.

## 3.4 Where to go from here?

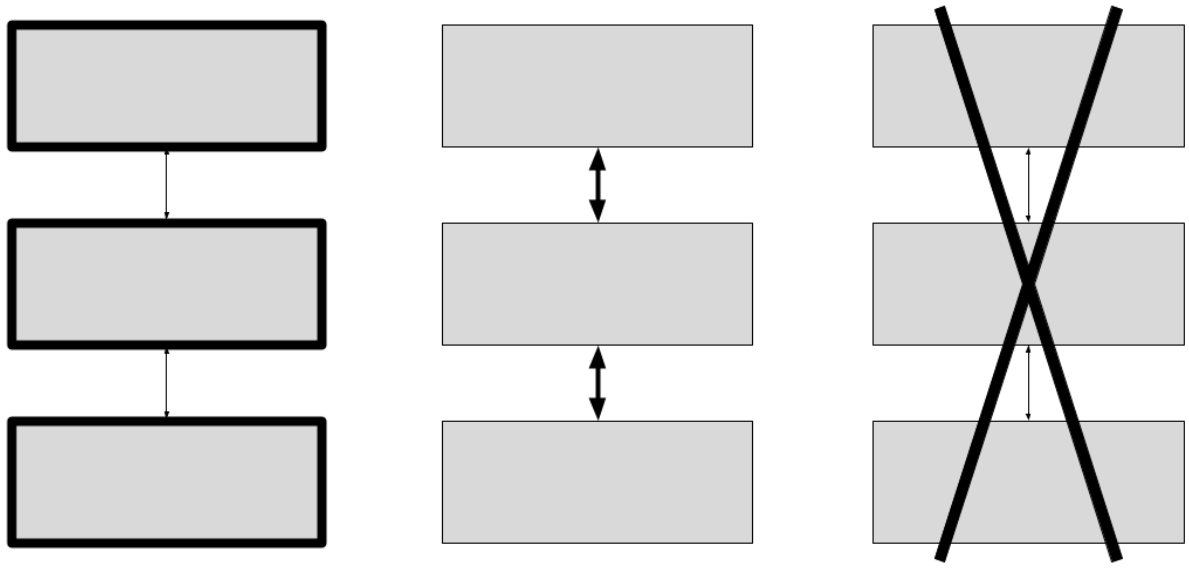


Figure 3.2: Three possible futures: a) Acting within the modules; b) Strengthening the interfaces; and c) Rejecting modularity.

Many efforts at supporting responsible AI, like AI fairness checklists [122, 165], model cards [184] and datasheets [98] assume panoptical visibility into the technology that our work demonstrates does not hold. Some have been designed as a kind of "nutrition label" [55], where facts are announced to an unspecified audience as if taking a view from nowhere. Other toolkits, such as Vallor et al. [255], acknowledge the interstitial nature of ethics failures, but when teams have neither visibility nor control over cascades of failure [i.e., 225], and do not believe they *should*, the success of inventory-like approaches is likely to be limited. If we instead start from an assumption of located accountability, where knowledge is partial and situated, we might seek places where there are relations between actors, and where people who are not developers have a

stronger role. While that is analytically straightforward in social scientific terms, it is more complicated for those who see the world through the lens of modularity, and who value the cutting of relations for specific reasons we have shown. Therefore, asking others to simply adopt located accountability wholesale will not do. We see three possible approaches, depending on how much our colleagues trained in the virtue of modularity are willing to question it.

### 3.4.1 Acting Within the Modules

If we fully accept that the dominance of modularity is unlikely to change soon, we would seek to act within it. Perhaps there is an opportunity for participants to *append* their partial understanding of the flaws, limitations, divergent provenances, and contexts of use of this documentation in checklists, model cards, and the like, thus relieving developers of the discomfort of being asked to definitively claim facts they felt they could not claim, as models and data changes hands.

This might require, ironically, doubling down on division of labor, by clearly delineating what knowledge on the card would come from developer's "module," and what comes from user experience, sales, and legal roles, leaving the supply chain metaphor largely intact. Nevertheless, this turns model cards into a boundary object where partiality comes together, even if deeper relations do not occur. This has obvious limitations. Unless there is a creative way to modularize participation from impacted groups, the very idea of which might be considered offensive, this approach re-inscribes their exclusion. It creates more modules for those who are not developers, but those include only what is publicly sayable. It leaves to regulators, journalists, and academics to force conversation and action about that which is considered unsayable from within the chain.

### 3.4.2 Strengthening the Interfaces

Another approach would move away from metaphors of supply chains towards a more managerial notion of "value chains," which orchestrate companies' activities in ways that that combine to create competitive advantages [85]. This would strengthen business connections between

companies beyond those allowed by the "developer hat," and buttress the communication that happens in the process of exchange. Model cards would be reinforced by contractual obligations and meaningful customer knowledge and communication, involving increased contribution from non-developers. Those in customer roles might scrutinize suppliers by asking for model cards, properly consented training data, and appropriate pay for data labelers, all scrutiny which is common in supply chains for physical goods. [134] propose technological measures strengthen module interfaces, by auditing AI services for misuse. These activities all help suppliers reframe ethics work as an act of delivering customer value. Still, this is not equally possible for every company. For example, one study showed some AI entrepreneurs conceal the ethics work they *were* doing from their venture capital funders interested in hiding limitations [271].

The interface between onstage and offstage would have to be strengthened too, to help people integrate their multiple locations in and out of the supply chain. Developers might leverage their value as difficult-to-find laborers by making clear they are not prepared to pay personal reputational costs, while journalists and academics could also place more emphasis on the multi-actor cascades [129]. If the supply chain centers on perfect control over one's module, a value chain might center on probabilities and frictions– what technologies, contractual obligations, or marketing messages make easier or harder, faster or slower. For example, the Ethical Source movement uses licenses to introduce legal friction for harmful uses in software supply chains, acknowledging this control is not total[2].

This approach facilitates the formation of stronger norms, bearing a surface relationship to values-sensitive design [94]. With numerous positionalities through the chain, "working misunderstandings" [88] where parties mutually misrecognize the actions of one another, are more likely than straighforward values alignment. Managerial notions of "value chains" often elide the problem of *who* value is created for, on the assumption that value is a function of what markets will pay for. Depending on policy conditions, this approach could risk setting up a path

[2]See: `https://ethicalsource.dev`

dependence where ethics issues can be better acknowledged and acted upon, but remain a second order, lagging concern where market value cannot be found.

### 3.4.3 Rejecting Modularity

What if modularity were eschewed entirely, both in terms of code and the broad social arrangements it mediates? Actors who object to the modularity ethos in the first place might abandon any notion of a chain entirely and prioritize building good relations as a matter of first order concern, building code second, and "scale" as a distant matter at most. Here, social relations cannot be bracketed off as a mere input or requirements capture. The relationship is the objective, not the lines of code that may or may not result. Any code that does develop might be in the service of questioning what software tools are necessary at all, and whether they need to be entirely different in different social conditions, as per Agre's critical technical practice [9]. Echoing criticism of endless AI scale [30], Gebru and Hanna propose such a model of AI development, where the goal is not to produce "AI for the value of AI itself", but to instead be "sensitive to other forms of knowledge" in order to examine and curate datasets even if this is slower or more expensive [244]. Here, differences between users *do* make a difference [see 113], while distinctions between producer and user begin to soften. One party is not the testbed for the other's "scale."

This approach might seem foreign to those building general purpose frameworks or scaleable "software as a service" architectures. Look just outside dominant norms, however, and there are plenty of examples to be found. Indigenous data sovereignty principles specifically call for exactly this kind of approach [51]. In her work with North Carolina community healthcare workers building vaccine equity for Black and Latinx communities, Gray [107] employed design justice principles from Costanza-Chock [65] to argue that "we must prioritize a deep, methodical connection with subject matter and domain expertise in lieu of an unexamined rush to scale or to shield ourselves from the realities of a social world." Gray recognizes that her two-year

intensive process introduced "friction, or working against scale, [which] is considered a bad thing in [Computer Science]. It is considered inefficient, a waste of engineering time." She recasts that ethos in the context of Arendt's banality of evil, and notes that frictionless "efficiency" is the very thing that creates a remoteness from reality, and opens the door to harm.

While the previous approach strengthened norms in a broad but inconsistent way, this does so in a more focused but deep way. Such focus has a long history outside an AI context [see 65, for an overview]. However, rejecting modularity in a modularized world raises interesting questions for upstream tools. Would someone fully reject all lines of code that were ever designed as modules in a literal way, or reject the broader belief system modularity entails and seek opportunities to build differently, or be more careful about choices of upstream components, like libraries or compilers, especially when made by companies known for ethics breaches? These choices might open up new avenues of technical innovation. In making them, teams might learn the specific ways that "generic" tools are not in fact generic at all, but generic only to those who are currently well served by the current supply chain. It might be that the need for other kinds of yet undeveloped "generic" tools that serve other interests becomes apparent. Finding and developing these would be a significant act of critical technical practice, and open up engineering paths otherwise foreclosed. This approach also raises questions for public policy. Given the resource inequalities between community groups and companies that seek to scale, and that those same groups are meeting social needs that arguably benefit a country as a whole, what would an appropriate science and technology policy do to support these efforts?

## 3.5 Conclusion

In this chapter, we have discussed how thinking about ethics and responsibility as chains of relations reveals specific locations in which ethical decision-making can take place. Those locations might be upstream or down, and they might be within the cultural logic of modularity or outside it. The combinations of these locations shape what is considered sayable and what is off-stage

talk. They shape what is prestige-garnering work, what is paperwork, and what is high stakes public drama. These social locations also shape the points of AI governance intervention, which rely on the extent to which actors themselves are willing to, and are capable of, acknowledging their own locations within a broader system of production, and engaging more fully in the relations in which they are involved. The core of the matter–how much modularized thinking should dominate software production–will not be settled easily. Consensus might not be achieved and multiple paths might be followed by different sets of actors with different visions of what responsibility is. Regardless of which directions others take, we have shown that realistic responsible AI interventions can start by making deliberate choices about how strong a role current software production ideals should play in future responsible AI development.

# Chapter 4

# Precarity, Powerlessness, and Workers' Ethical Concerns

Work in this chapter was originally peer-reviewed and published at the 2023 *ACM Conference on Fairness, Accountability, and Transparency*, with coauthors Derrick Zhen, Laura Dabbish, and James Herbsleb [269].

## 4.1 Introduction and Related Work

Facing public pressure and negative press [116], many large technology companies are attempting to address harms from algorithmic systems, often by instituting ethics initiatives which converge on principles such as transparency or fairness [137]. Metcalf et. al show how broad Silicon Valley logics cloud official ethics initiatives [179], some startup environments see ethics work as premature [254], and even some "major companies" see ethics work as "too complicated for the organization's current level of resources" [216]. A variety of interventions have been proposed to make operationalizing AI ethics easier, including fairness checklists [165], fairness toolkits [277] datasheets [98] and model cards [184]. However, some argue that the convergence around codified principles like "fairness" or "accountablity" obscures underlying political and

normative disagreements [185], and there is increasing evidence for this: AI practitioners have different values than the general public in AI system design [130], and workers have different concerns than those who seek to monitor them [203]. These principles may also have discordant definitions [79, 86, 161], and others argue that principles limit scrutiny to a system's design without scrutinizing *use* [142] or business decisions [108]. These concerns lead to accusations of "ethics washing" [114, 259]: where companies put forward voluntary principles to burnish their reputation and avoid regulation [196], without changing their behavior [279].

Given that software practitioners have some agency in how to develop these systems [201], research examines their needs and behavior as they seek to build ethical systems. For example, past work questions the effect of ethical codes [104] on software engineers' ethical decision making [177]. In machine learning specifically, Holstein et. al examine practitioners' challenges in developing fair systems, Madaio et. al examine practitioners' challenges in using disaggregated evaluations to assess system fairness [164], and Veale et. al examine the needs of public sector practitioners in ensuring fairness and accountability in high stakes systems [258]. However, as discussed above, studies which focus on "fairness" or "accountablity" may impose a narrow scope of scrutiny and thus foreclose on wider concerns, and many software practitioners work at smaller companies that may not have official ethics initiatives. Given these concerns, we surveyed 115 and interviewed 21 software engineers about their *self-identified* ethical concerns, as opposed to concerns identified using codified ethical principles, toolkits, or codes, to answer: **RQ1: What are software engineers' ethical concerns?** With this open scope, we discuss both the kinds of concerns our participants raise – military, privacy, advertising, surveillance, and others – but also examine the scope of their concerns: ranging from concerns about bugs which can me more easily fixed, to wider concerns questioning their company's *raison d'être*.

Others study what happens after tech workers of various stripes develop ethical concerns. Whereas some AI practitioners engage in high-profile activism [29], Madaio et. al find that others advocate less strongly for fairness issues due to career concerns [165], and Richmond Wong

shows how User Experience practitioners employ softer tactics of resistance [276]. Others study how relatively less powerful gig workers resist opaque algorithmic evaluation [215] and use online forums to seek to understand algorithmic management they are subjected to [157], and others study how crowd workers engage in collective action [224]. Nedzhvetskaya and Tan collected examples of blue and white collar tech worker engaging in collective action [191], and discuss how workers claim they ought to have a role in AI ethics governance [190]. Similarly, after collecting software engineer's ethical concerns, we investigate how they respond: **RQ2: What happens when software engineers develop ethical concerns?** We report on a broad variety of actions participants take – from proposing technical fixes, to negotiating within organizational incentives, to resigning in protest – and on the psychological toll that these actions lead to.

Within tech ethics research, power is increasingly recognized as a central factor when differently situated actors raise concerns. For example, recent work examines power asymmetries as students resist algorithmically lowered grades [31], and how software engineers see themselves as less powerful "mediators between powerful bodies" [201]. Others position software engineers as powerful actors in AI ethics given high demand for their labor [58]. A recent critical analysis of AI fairness toolkits find that they frequently ignore organizational power dynamics [277], and a recent review study finds that future work ought to attend to "structural and historical power asymmetries" [38]. In line with this call, we examine contingencies of software engineers' power as they raise ethics concerns, with an eye towards how these contingencies explicitly factor into the actions they choose to take: **RQ3: What affects software engineers' power to resolve their concerns?** We find that financial and immigration precarity, workplace culture, and organizational incentives constrain participants' power to see their concerns resolved.

After detailing our survey and interviewing methods, successive sections answer each research question. In the discussion, we ground our analysis of how power affects practitioners' ability to raise ethical concerns in frameworks of power from Organization Science. In particular, we draw on Fleming and Spicer's framework of power in organizations [91], which is composed

of *episodic* "faces" of power, where managers give orders directly or seek to limit discussion to within acceptable boundaries, and *systemic* faces of power enacted by constructing hegemonic ideological values within an organization or by seeking to shape practitioners' sense of self. We draw secondarily from other frameworks of power from organizational science when examining how practitioners are disempowered when they face paradoxical demands at work [33], and when discussing how practitioners may exert agency to resist unethical assignments [152]. We then discuss the implications of our work: that future tech ethics research ought to turn from helping spot issues to helping practitioners build their power to actually fix them, and we also question the foci on AI or Big Tech in tech ethics discourse.

## 4.2 Methods and Participants

We seek to understand practitioners' self-identified ethical concerns and how they navigate them. Therefore, we imposed no *a-priori* definition of ethics, nor do we seek to reach a singular definition in our work: instead, in our survey instrument, we use an open-ended framing to ask survey respondents if they have "ever had ethical concerns with a software system they were asked to contribute to" and actions they took, resolutions, factors which made raising concerns harder or easier, and an invitation to an optional follow up interview. To recruit a broad sample, use recruited using diverse methods including posts to Twitter, software engineering message boards, software-ethics focused messaging channels, the popular StackOverflow programming Q&A site's blog; and in person at a developer meetup.

The survey was open for 87 days from May to August 2022, and received 115 responses. 90 survey respondents were employed full-time, 15 were employed part-time or as contractors and 10 were not currently employed. 13 respondents worked at very small firms (¡10 employees), 29 respondents at small firms (10-99), 31 at medium sized firms (100-999) and 35 at large firms (1000+), 7 did not report the size of their firms. Respondents were relatively experienced, reporting a mean of 17 years of experience coding (med. = 15, min. = 4, max. = 46 years).

Respondents spanned six continents: 68 participants lived in North America, 34 in Europe, 4 in Australia, 4 in Asia, 3 in South America, and 2 in Africa. 80 participants identified as male, 10 as female, 6 as nonbinary or nonconforming, 5 self-described and 14 preferred not to answer. 21 survey respondents participated in the optional follow-up interview (demographics in Table 4.1).

We conducted semi-structured [264] teleconference interviews to collect a detailed order of events as practitioners navigated their concerns, to probe into the their recollections of their thoughts and feelings, about factors affecting their agency and power to see their concerns resolved, and their work since. Interviews were recorded with participant consent and IRB approval and lasted between 21 and 73 minutes (mean, med: 41 min.).

We analyzed survey and interview responses sequentially. The first two authors performed an open qualitative card sort [284] on survey responses, negotiating disagreements and adjusting categories as necessary. On interview transcripts, the first two authors performed two rounds of iterative [262] thematic analysis on this data [43]: an initial round of open coding, and then the development and application of a closed coding frame. Our study *makes use of* self-selection [235] to recruit those with self-identified ethical concerns without any pre-ordained scope, but therefore our results *do not* support general claims, such as the overall prevalence of a given concern. Interviews and surveys collect self-reported experiences, risking social desirability [189], and hindsight biases [118]. Instruments were in English, a widely-spoken language for intercultural engineering communication [218], but our findings may not generalize to software engineers working in other languages.

## 4.3 RQ1: What are software engineers' ethical concerns?

We answer this research question in two ways: firstly, explaining the *kinds* of ethical concerns raised in our survey most frequently as surfaced by our card sort. Secondly, as a spectrum illustrating the different *scopes* of practitioners' concerns, according to how much of their organization's priorities their concern calls into question.

### 4.3.1 Kinds of Ethical Concerns

#### Military

17 practitioners wrote about concerns related to military applications of their work. Of practitioners who disclosed details about the systems they worked on, the most common concerning system was autonomous drone navigation software (n=5): *"Work on autonomous drone visual navigation in a GPS-denied environment" (S98)*. Other respondents develop training software: *"software in support of simulations used to train US warfighters" (S161)*, logistics software for military organizations:*"I contributed to the development of a proprietary platform-as-a-service used in defense contracts" (S174)* and engineering support software: *"an analysis tool that automatically finds errors in aeroplane jet engines." (S188)* Respondents were primarily concerned that their work would physically injure or kill others: *"I was concerned whether the software I was contributing to was being used to harm innocent civilians or infringe on human rights" (S174)*, but several raised broader ideological concerns with the militaries who used their systems, one asking: *"am I indirectly contributing to the ills of imperialism?" (S161)*

#### Privacy

14 practitioners expressed concerns relating to privacy, most commonly about geotracking (n=4), one saying they *"grab[ed] geolocation data from customers [but] our product doesn't use geolocations." (S67)*. Others were concerned about *"stor[ing] user keystrokes in a signup form to a marketing and analytics platform before the user actually submitted the form" (S272)*, scraping social media profiles *"as part of additional information to include when making loan decision." (S114)*, and privacy involved in data labeling on private footage: *"contractors [were] to label hundreds of thousands of [home security] video clips" (S26)*, or requiring personal data *"not necessary to have for the task at hand" (S69)*.

### Advertising

13 survey respondents reported concerns about advertising. Practitioners were concerned about building spam email systems, or *"bypass [spam] prevention measures" (S139)*, concerned that spam breached customers' privacy (S86), delegitimized email marketing (S139) and did not do good in the world (S230). One respondent wrote that being asked to *"develop a computer vision system that accurately classifies someone's demographics for customer segmentation marketing" (S78)* as something he believed to be inherently racist and sexist. Other practitioners wrote about implementing dishonest interfaces to *"push users to buy something because stock was "almost out"" (S2)* when in fact it was not, helping to air ads that were*" degrading toward women" (S22)*, and about advertising *"scummy for-profit schools." (S102)*

### Surveillance

11 respondents described being asked to contribute to systems used to surveil workers or citizens. Four respondents recounted concerns about working on existing workplace surveillance and algorithmic management (i.e. [157]) systems, such as *"observing how well grocery stockers stayed on task" (S82)*. Their concerns included *"overwork [and] anxiety" (S82)*, that it might be *"illegal to measure employees' pee time" (P74)*, and that *"low sales numbers" (m)*ight be used to unjustly*"fire employee[s]." (S10)* Other practitioners were invited to work on surveillance systems for governments. One interviewee (I14) was asked to architect an intelligence gathering platform for a foreign government. Another respondent made improvements to an existing telecom surveillance system (S13). Other practitioners did not build surveillance systems directly, but were worried their system might be used as such downstream: *"the big problem was that I didn't see a way or a use case, where [facial] identification would be used in a non-ethically problematic way. So those would be at frontiers, at airports, identification in police stations." (I14)*

**Environment, Labor Displacement, Inequality, and others**

Categories of concerns expressed by less than 10 practitioners included environmental impact (n=4) *"monitoring system for agropecuary [livestock] business [which] is highly damaging to the environment" (S21)*, labor displacement (n=3) *"I thought the software system could very well put some people out of a job" (S201)*, and exacerbating inequality (n=3) *"statistically, there's no way they could do this without some form of systemic discrimination." (S44)* Other harms cited included overcharging customers (S54), contributing to addicting products (S150, S174), cryptocurrency as multi-level marketing (S70), inaccessibility of software (S50), jeopardizing healthcare outcomes (S66), legality (S115, S95), botnets (S133), implementing dark patterns (i.e. [106]) (S100), autonomous vehicle safety (S118), and political manipulation (S104). Some had concerns with the software development process itself: using vulnerable frameworks (S143, S39), underpaid data labelers (S26), or closed-source software (S106).

### 4.3.2 Scope of concern: concerned with a bug, or your whole industry?

We also found that ethical concerns varied wildly in *scope*: varying in how much the organization's goals or priorities a given ethical concern questions. While they overlap, we illustrate this using four scopes of concern: those arising from bugs, intentional features, whole products, and finally concerns which question their organization's *raison d'être*. Scope affected outcomes: concerns questioning entrenched organizational goals were harder to resolve (see Sec. 4.5.3), and affected the kinds actions practitioners took (Sec. 4.4).

**Bugs**

Some practitioners described fixing bugs as their core ethical obligation, one saying: *"for a software developer, [software] quality is the core of ethics. Because if your product is unreliable, then your representations about the product are probably unethical." (I17)*. In some cases, proposing to fix bugs is uncontroversial, since maintaining intended functionality is often within

an organization's best interest. For example, when a practitioner raised concerns about a bug in construction crane safety, they the practitioner described how this was enthusiastically received and resolved: *"there were a lot of really high profile accidents with lifting cranes [...] Everybody was really super on edge about making sure that our simulations were correct. [...] And so when I brought that issue up [...] they did a big investigation and found out that it was a data entry error." (I10)* However, organizational incentives can instead stifle practitioners' efforts to identify, fix and prevent bugs. For instance, one respondent felt non-technical firms tend not to invest in code maintenance as long as the software is minimally functional: *"non-tech companies [...] just care about business continuity" (S81)* Another interviewee explained how cost cutting at his consulting firm made it difficult to do work of acceptable quality.

**A specific feature**

Unlike bugs, features were intentional: practitioners were directed to implement them by their manager or client, and therefore questioning them often required more directly questioning their organizations' objectives. For example, one interviewee was asked to implement a feature that would round down GPS coordinates on properties being evaluated for insurance eligibility, which *"would have denied people access to certain types of insurance." (I10)* Another interviewee working on workplace compliance software reported that his boss asked him to implement a feature that he felt was privacy invasive: *"My boss [said] we need to put in a thing on the app so that we can see where people are all the time. And I told him [...] most of the people install it on their personal phone." (I6)* Other concerns arose when practitioners were disallowed from implementing features they felt were ethically important. For example, an interviewee developed ethical concerns about how her product may be exclusionary: *"A really famous VR software at the time, had done inclusivity in terms of the color of the skin [...] and allowing for people with one hand to operate it. [...] I brought it up as an option" (I11)*, but this was not pursued and she was told *"well, nobody asked for it." (I11)*

### An entire product

Practitioners also surfaced ethical concerns about entire products, or, as consultants, entire contracts they were assigned to. When respondents had concerns about a product's very existence, many felt concern could only be resolved if the product is shut down or dramatically altered. One contractor at a marketing consulting firm was assigned to develop a customer segmentation model, to help their client profit from high interest loans by: *"find[ing] customers that were likely to [...] take on unsustainable amounts of debt." (I5)* In this case, changes to the implementation of the product would not reconcile the practitioner's concern that building a product to sell "unsustainable" loans was unethical. Another interviewee reported being assigned onto a project to make improvements on telecom software which he suspected was being used for telecom surveillance: *"One of the main managers mentioned that the their main client for the device at the time was AT&T. [...] based on what the device was doing, they figured [...] the main use case [was] NSA tracking." (I13)* In this case, the practitioner's concern was with misuse of the product he was working on, which could not be resolved until the product was terminated, or its core use cases rethought.

### An organization's *raison d'être*

Finally, some practitioners reported concerns with their organization's or industry's goals or business practices. Many practitioners were concerned that their work was used for military purposes, constituting the most common concern type. These included concerns of direct harm, such as *"the software I was contributing to was being used to harm innocent civilians" (S174)*, but also ideological issues, one pondering *"am I indirectly contributing to the ills of imperialism?" (S161)* One practitioner cited his newly-held Buddhist faith as the origin of his concerns that working in the "weapons domain" at all is *"really not good karmically" (I1)*, later reflecting that *"if you pay attention to what was going on, like in the wars, it doesn't have to be so esoteric as like Buddhist precepts." (I1)*

One interviewee, working at a fintech firm, felt his work *"preventing [fraudulent use] was not really an ethical challenge. The issue was more than the company as a whole, the business model [...] It was, you know, payday lending." (I2)* In this case, the interviewee felt was concerned about the very reason the company existed, reflecting that this made raising any concerns feel futile: *"you're actually asking to shut down the business. [...] you might as well say to the founders, like, 'Hey, either you shut down or I'm leaving', and they'll be like, 'Alright, leave, I guess.' It's not really a concern you can raise." (I2)* Even firms that offer services instead of products can be held to this level of scrutiny; as one practitioner held that their consulting firm's willingness to do business with shady clients comprised a core part of their business model: *"The company [...] does a fair amount of work for [...] oil companies, [...] firearms, [...] British American Tobacco [...] not exactly paragons of morality." (I5).*

## 4.4 RQ2: What happens when software engineers develop ethical concerns?

### 4.4.1 Technical Solutions

Some practitioners proposed technical solutions — changes in the functionality or design of a system through code modifications — in an attempt to mitigate potential harms. Technical solutions work best on *Bug* and *Feature*-scoped concerns, because harms resulting from the core purpose of a product or the business practices of an organization (i.e. those later in Sec. 4.3.2) are not able to be resolved through changes to system implementation.

Furthermore, even when practitioners see opportunities for technical solutions, their actual implementation depends on management agreeing that perceived harms are important enough to warrant dedicating resources to fix them. For example, both interviewees I6 and I10 (whose concerns were summarized in 4.3.2) came up with technical solutions that would have resolved their concerns, which were dismissed by management. Interviewee I6 came up with a design

affordance to minimize the privacy concern he had about employee location tracking: *"if you really desperately wanted to [...] see where each person is on site, [...] we could geofence the site [...] If they're not in [the site], the tracking is off." (I6)*. While management was sympathetic to his privacy concerns – *"[my manager] agreed [...] we cannot monitor people's comings and goings" (I6)*, his geofencing solution was ultimately rejected due to resource constraints: *"he blatantly told me that's too much work. And he's not signing off on that." (I6)*. Interviewee I10 proposed a solution to avoid erroneously denying people insurance coverage due to GPS rounding errors, suggesting *"we [could] have a three value response [the third being] 'maybe need to check further if it was right on the boundary"' (I10)*. However, what the practitioner had experienced as a serious concern *"people need flood insurance for their houses [...] I had been victim to flooding and lost a bunch of my stuff" (I10)* was a non-issue for the client: *"the client cut me off and told me she didn't care and that [...] I just needed to do it." (I10)* He was later *"dressed [...] down for speaking out of turn with the client" (I10)*, and the manager *"threatened to fire me if I didn't do the work." (I10)*

### 4.4.2 Negotiating within organizational incentives

Practitioners also sought to resolve ethical concerns by convincing decision-makers like engineering or product managers that harms are serious enough to warrant action. Often times, this involves phrasing ethical concerns in terms of their effects on organizational incentives such as profit or product success.

For example, one ML researcher concerned about his project's use of facial identification (i.e. who is in this picture?) reported successfully pivoting the direction of his project to facial verification (i.e. are these two pictures the same person?). He raised ethical concerns about downstream uses like bias and surveillance to his management, but couched these within organizational incentives to pursue an easier and more achievable project (verification) instead of a more difficult one (identification): *"because we were understaffed [I said] ' [...] we don't have*

*the resources to do it."' (I14)* Another interviewee, who had ethical concerns about improper employer vetting in a job matching application he helped develop, described attempting to get senior management to shutter the project by appealing to the organization's core values. *"[I said] we either need to invest more money into understanding what is going on here [...] or we need to pump the brakes [...] I was quoting, you know, our organization's code of ethics and stuff like that." (I9)*

The likelihood of ethics negotiations succeeding are, as one practitioner puts it, *"entirely [dependent] on the organization and your ability to talk to people and [...] capture hearts and minds." (I2)* A practitioner's ability to affect change internally through "rocking the boat" relates to the broader work Debra Meyerson has done on "tempered radicals" [182] — leaders who leverage their status within organizations to promote their own values and ideals. The approach of affecting change from atop the corporate ladder was also suggested by one of our interviewees: *"[you could] work your way into a leadership position, and then start making different kinds of ideas" (I5).* However, they acknowledged the fraught existence of individuals attempting both conformity and rebellion: *"you'd have to both hold on to your ideals [...] And at the same time, be willing to compromise your ideals quite heavily in order to work your way into a leadership position in the first place." (I5)*

### 4.4.3 Refusal

One common action respondents reported was refusing to work on the task they found unethical. Refusals took on various forms, the first being "quiet quitting" – reducing one's productivity. One practitioner who was asked to build a system to bypass spam filters wrote: *"I purposefully created a poor implementation and did not dedicate very much energy to make a working solution." (S49).* Another respondent wrote that they *"pretended to complete the task but didn't" (S62).* We found that the tactic of "quiet quitting" emerged from a feeling of powerlessness to affect change within organizations, and as a result is often accompanied by searching for other

jobs (see 4.4.4). One practitioner who reported reducing productivity felt that it was impossible to resolve their concerns internally, since the product they were concerned about was already in production: *"I don't think I had any power in this dynamic because [the product] was already deployed. This was just like a minor upgrades [to] make it more usable." (I13)* Since the practitioner saw little utility in pursuing a resolution internally, they *"reduced productivity to a minimum and found another job" (S6).*

A handful of respondents reported seeking reassignment to a different project. These practitioners removed themselves personally from the concerning project, but did not attempt to use their leverage to shut the project down: *"I was given another project to work on. I didn't kill the project, but I also didn't contribute to it." (S19).* Reassignment is typically only possible at organizations with many product lines or clients, and practitioners felt they needed seniority to ask for reassignment, as one described: *"My seniority and wide swathe of other projects to choose from" (S19)* made securing a reassignment easier. One participant described a policy to make it easy to seek reassignment on ethical grounds: *"We had a policy at the company that nobody has to take part in any software projects involving military use." (S95)*

Other practitioners delivered ultimatums to management – putting their job on the line and making it clear that they would quit unless their concern was addressed, with mixed results dependent on their leverage and the scope of their concerns. One practitioner working on hospital software was concerned that the rushed rollout of an update would jeopardize patient outcomes. In response, he raised the concern to management forcefully: *"I looked that manager in the eye and I said: you are going to have to write me up or fire me, but I'm not doing it. I'm not going to put patients' lives at risk, because you've got a pile of money sitting on the table." (I12).* This confrontation resulted in management stepping back and reassessing the necessity of the update. One participant suggested that ultimatums can be a wake up call for management, forcing them to take seriously harms they may have ignored in the past: *"maybe [leadership] didn't know how the individuals in the org felt. And then, individuals in the org might raise a stink. And sometimes*

*that leads to some work being paused or just like not being done." (I7)* However, they suggest that the effectiveness of an ultimatum is highly dependent on how much leverage a practitioner has, and that collective ultimatums tend to be more effective: *"if it looks like we're gonna lose a big chunk of employees, [management] might say, we can't afford that [...] it kind of depends on the individual, whether you have leverage over leadership." (I7).*

Resignation is typically a last resort: practitioners resign after their technical solutions or compromises are rejected (I6, I10); when escalations go sour: *"he threatened to fire me if I didn't do the work. And that's when I decided I would just quit." (I10)*; or when they lose faith that the ethical concern can be resolved internally: *"Raised concerns with executives. Started ethics discussion group among employees. Left the company after seeing no progress." (S53).* Resignation allowed participants to put distance between themselves and the projects they deemed harmful, but they often reported this as bittersweet: in resigning, they relinquish control over development of the harmful system, as another developer is often hired on and progress resumes. One survey respondent lamented this, saying his concerns were not resolved because *"the company hired someone else. [...] I felt that I would have been in a better position ethically if I had taken the contract and had done a bad job of it." (S79)* However, in some cases, the resignation of a crucial developer in an already precarious project can terminate the project. One participant reasoned that their departure likely doomed the project: *"I was also the only one who had any serious level of software development competence [...] they generally struggled with deploying the existing models [...] so I can't imagine that they would have deployed it." (I5)* In another instance, a contract worker heard that his client canceled the project he worked on after his resignation, reflecting: *"[... quitting] can give the client cold feet on the project, it makes it look like the consulting firm is incapable of managing the project. So [...] they're likely to just cancel the project completely." (I6)*

### 4.4.4 Feet voting: "This work doesn't get done without us"

The strategy of "feet voting" describes the proactive actions practitioners took to align their employment decisions with their ethical views (such as career planning), in contrast to reactively refusing assignments or quitting jobs due to an unresolved concern. The most common action reported in this category was refusing offers of employment. Either turning down a job offer: *"I rejected the offer" (S47)*, dropping out of the interview pipeline: *"I decided to not continue interviewing with said job" (S45)*, or deciding not to apply to a position: *"Ignore the job advert" (S82)*. Many saw turning down employment to be easier than resigning, but others lamented passing up lucrative jobs: *"[Anything that made it feel harder to act?] Just the big bag of money." (I17)* or interesting projects: *"I love game development, but I don't like to work for a company that does business in gambling." (I150)*

Some practitioners with concerns about their previous industrys' *raison d'être* went to great lengths to transition to another industry. But past experience makes this difficult, as one participant trying to transition away from developing war-fighting simulations said: *"It's difficult because my experience in this industry makes me most attractive to other companies working in the same industry." (S88)* A different practitioner found it necessary to move to an entirely different state to find opportunities he was ethically aligned with: *"I realized, well, if I'm going to stay in this area, like the odds of me at some point, working [...] on defense contracts are pretty high. [...] I'm being kind of a picky applicant on what companies I'll work for. And if I really want to do that, then I might have to consider moving [...]" (I1)* He also described being more intentional in screening potential employers for red flags: *"I realized you really have to look at like the ethics of the corporation, like, as part of your interviewing process [for example, in the interview] I just asked about the details of the project [...] what space they were in, what type of product they were selling, that sort of thing." (I1)*

One practitioner argued that the favorable software engineering job market implies a unique ethical responsibility: *"Even like the 2008 financial crash [...] every software developer I knew*

*still had work. Even if the job they had disappeared, they had a new one within a week or two. [...] I think software development is incredibly resilient against recession [...] that's why we have a responsibility to be sticks in the mud about ethics. This work doesn't get done without without us." (I10)* However, more often, participants did not feel this way (see 4.5.1).

Collective bargaining and tech worker boycotts are instances of feet voting at scale, in which practitioners collectively withhold labor from organizations they had ethical concerns with. These tactics have grown in prominence at large tech firms [146]. However, among the practitioners in our study, only one interviewee raised this *"the company would have to be pushed and that'd have to be either externally through [...] legislation or similar tools, or just public opprobrium or internally through unionization" (I5)*, mentioning that *"I did attempt to do a bit of [union] organizing work. But unfortunately, I was doing that alone." (I5)*

### 4.4.5 Leveraging legal systems

One practitioner we spoke to attempted to collect information to raise his concern with law enforcement but did not ultimately go through with it: *"I knew [...] they were going to have to start skirting rules right from the start. So, so yeah, I asked for all of the requirements, documents, anything you could give me to help me understand how to build such a system [...] My intention was just to walk into the FBI." (I17)* Another interviewee echoed this idea, saying that for harms that call into question the *raison d'être* of the entire organization (see Section 4.5.3), external enforcement was the sole option: *"if you do have a concern, you should take it up with the legislators or the courts." (I2)*

Practitioners who maintain open source software can also leverage laws around software licensing to prevent misuse. For instance, one practitioner personally opted to use a "copyleft" (i.e. see Chapter 2.4.1) license in order to limit downstream harms of OS agricultural software they created, but conceded that it was unlikely that they would have the resources for costly litigation to enforce them. In discussing the efficacy of their action, they compared the process

of choosing a license to what they saw as the small and easy yet important effect of voting as a way to effect change: *"it's a small, little one time thing you can do, that probably won't help you. But, but if it does help you, it is huge. And it only took two minutes of your time to set in place, and it's there for years, you know, that you may need to fall back on that if that's your only line of defense." (I19)*

### 4.4.6 The psychological toll of raising concerns

Practitioners reported experiencing anxiety, depression and isolation throughout the process of identifying and raising ethical concerns. The process of raising ethical concerns to an employer was stressful, especially for full-time employees, for whom their organizations are their sole benefactors. One practitioner writes: *"it terrified me to confront an 'authority' figure, especially one who was the source of my financial well-being." (S62)* Another practitioner described raising ethical concerns with a client as: *"one of the most terrifying moments in my life." (I10)* The aftermath of a failed escalation can also seriously affect practitioners' mental health, as one interviewee recalled: *"I spent a good few weeks lying in my bed [with] serious depression [...] I didn't want to leave my apartment [... I] just couldn't face [...] checking work emails." (I5)* After his concerns were dismissed by both the client and his direct manager, another described *"It gave me a lot of anxiety and depression. [...] And it kind of made me cynical [... I] approached most new working situations [...] trying to not get too involved [... ]just so that it would be easier to cut and run, if somebody asked me to do something unethical." (I10)*

Practitioners also reported that just having an ethical concern at all was distressing. One interviewee quit multiple jobs over ethical concerns, recounting *"I was so distraught over what I was being asked to do, I threw up in the parking lot before going into work." (I10)* Another interviewee spoke about the alienating effect of being the only person in the office with an ethical concern: *"[I felt] kind of like an outcast" (I1)*, and another survey respondent suggested that raising concerns could lead to hostility: *"I do not want to judge, or be judged, by colleagues*

*for my views. Without care, such discussions can lead to a hostile work environment." (S38)* However, others circulated concerns among peers in order to feel less isolated in their concerns. One interviewee leveraged their organization's employee directory and intranet to *"find other people who cared about the same things" (I4)* and start ethics reading and discussion groups. Looking back, they reflected: *"Finding community in the ethical AI space made me feel so much more grounded." (S14)*

## 4.5 RQ3: What affects software engineers' ability to resolve their concerns?

In this section, we discuss personal and organizational factors which affect practitioners' ability to see their concerns satisfactorily resolved, including financial and immigration precarity, company culture, and organizational incentives.

### 4.5.1 Financial and Immigration Precarity

While some software engineers felt comfortable turning down jobs (Sec. 4.4.4) or quitting their current jobs (Sec. 4.4.3) over ethical concerns, many practitioners expressed financial limitations on their power to act on their concerns. One explained how concerns over precarity took priority over ethics: *"Any kind of precarity will make your weigh your ethics less, right? [...] having a family, having dependents who can't support themselves, [...] medical conditions [and given this] you kind of are able to talk yourself into, hey, [...] I don't really have a choice." (I2)* When asked about anything that made it harder to act, survey respondents echoed this: *"The need to provide a living for me and my family, together with high prices" (S82)*, *"Reliance on the job to survive" (S8)*, and simply: *"Money." (S77)* Survey respondents also cited financial stability making it easier to act: *"I was single, didn't have a lot of debt" (S20)*,*"I had a decent savings and could afford to drop the client." (S62)* One interviewee described a stark example: *"aside*

*from [the ethical concern...] my father had passed, and so I got some life insurance money [...] so I didn't necessarily need the paycheck anymore." (I21)* Support networks mitigate precarity: *"[My parents] said [...] they would help [me not] get put out on the street." (I10)*, but so does lacking dependents to support: *"I'm only supporting myself." (I11)*

Precarity from employment-based immigration visas (e.g. US H1B visas [101]) also influenced whether practitioners decide to take action, one interviewee making clear he would only ever leave a job if he had another opportunity lined up, saying his semi-permanent state of precarity leaves immigrants less freedom act on their ethical scruples: *"Indians on H1Bs [often] need to find something [a job] within a very short period of time or actually have to leave the country. And when that happens, you end up taking whatever is available." (I2)* Practitioners were also worried about blacklisting, as one stated fear over *"getting [...] bad recommendations from former employers." (I10)* One interviewee described being blacklisted after raising an ethics concern: *"[the director] sort of ended it with like [...] I can't fire you. Because you're in contract. But like, know this: the aid sector is small. And your career here is like pretty much over." (I9)*

### 4.5.2 Workplace Culture

Respondents described how their organization's culture – including norms, expected practices, and communication styles – affected their willingness to raise concerns. For example, participants cited *"trust and respect [and] a common goal" (S71)* and an *"'Open door' policy [...] easy to get 1-on-1 time with execs" (S52)* as things that made it easier to act on concerns. However, more respondents described *"hostile" (S72)*, *"authoritarian/passive aggressive management style [...] hierarchical culture" (S110)*, *"suggestions from higher-ups that ethics discussions were a waste of time." (S52)* as things making it more difficult to act on their concerns. Interviewees expanded on this, one (I10) contrasted his two consulting experiences: the first where he worked in cubicles *"in a building full of thousands of people and feel lonely" (I10)*,

and the second where he *"had good rapport"* and *"trusted"* his client and therefore felt *"safe"* to bring up his concerns.

However, some participants said "friendly" cultures made it *harder* to raise concerns. For example, survey respondents recalled that because *"The boss was a friendly chap" (S104)* or *"bonding attempts from the owners" (S68)* made it harder to raise concerns. One interviewee said that remote work meant fewer social ties, which made it easier to escalate his concerns: *"This remote way of working [...] helped me to create [...] this disconnection with the manager [which] helped me to say [...] I care less about your opinion on this." (I14)*

### 4.5.3 Organizational Incentives

Participants demonstrated an acute awareness of organizational incentives, and used them to reason about their power to act on their ethical concerns. Profit motives lead to ethical concerns, as many survey respondents identified explicitly: *"features were implemented to earn money by any means necessary" (S69)* or *"they were selling geolocation data because it's worth a lot of money." (S67)* One interviewee said that financial struggles lead to "uncomfortable" tradeoffs: *"between the choice of closing the business [versus] doing something uncomfortable, almost everyone chooses to do something uncomfortable," (I7)* recalling that a previous employer sold user data to advertisers when *"scrambling [to find] some new revenue stream?" (I7)* As seen in Section 4.4.2, practitioners couched ethical concerns within organizational incentives to gain support. Consequently, one interviewee described how "ethics wins" were not about an *"ethical concern, but a marketing concern, to be honest. And the way that incentives align." (I2)*

Other practitioners suggested that it is easier to resolve ethics concerns at government agencies and nonprofits, as one interviewee who had recently transitioned into public service described: *"you're pursuing your goods beyond profit, right? [...] versus 'we want to make money'." (I5)* One interviewee doing software engineering at a public university described how state funding shaped project priorities, at least in the ideal: *"If the companies paid us, I guess*

*the situation would have been a bit different. But we [wanted] to work in the best interest of the people [...] we are paid by the people's tax money." (I20)*, but another academic described how pressures to publish led to his concerns about research integrity.

Multiple contractors and consultants described needing to compromise ethics to appease clients. One interviewee who resigned from a consulting project after being asked to do something illegal said: *"Your interactions with the client weigh very heavily on future decisions for future engagements and contracts. So there's a lot of pressure [...] to get along with the client. [...] If your client asks you to do something you don't want to do, too bad." (I12)* This appeared especially pronounced at financially precarious firms who felt the need to *"act on clients' whim[s]" (I71)* , or non-profits who feel accountable to donors rather than beneficiaries, where donors may instead have less beneficent geopolitical interests: *"the goals of the [project], are largely to keep refugees [...] in the Middle East. So they don't affect people in Europe." (I9)*

## 4.6 Discussion: It's not about spotting issues, it's about having *power* to resolve them

Identifying these concerns is only half of the struggle, and an unfulfilling one without the ability to ensure they are resolved. Despite recent layoffs [139], software engineers are relatively highly paid, mobile, in-demand and therefore relatively *powerful* [58] – yet our work shows that power is still a limiting factor in our participants' ability to ensure their concerns are resolved. In this section, discuss the centrality of power in raising and resolving ethical concerns, and implications for future tech ethics research, interventions, education, and activism.

### 4.6.1 Putting practitioners' power under an Organization Science lens

We have seen a variety of ways that participants sought to exert power as they raise ethical concerns (see Sec. 4.4), and factors which often limit their ability to do so (see Sec. 4.5). Here, we

discuss our results in light of frameworks of power in Organization Science, drawing primarily on that of Fleming and Spicer [91], who outline a framework containing discrete, *episodic* "faces" of power including *coercion* and *manipulation*, and *systemic* faces of power including *domination* and *subjectification*. They also note four "sites" of power, including power exercised *in* organizations' boundaries, *through* organizations as a vehicle for wider change, ways that external "elites" exert power *over* organizations, and outside power struggles organized *against* organizations. We also secondarily draw on Berti and Simpson [33], who use Fleming and Spicer's framework to understand worker disempowerment. Lawrence and Buchanan [153] also prove useful for their attention to how actors exert agency within their institution, even though their framework and language overlap yet is not commensurable with the other two frameworks, risking confusion.

In many cases, the forms of power enacted on practitioners were overt and direct, of the form Fleming and Spicer call *coercion*: being simply "told what to do 'or else"' [91]. We see this in how practioners' proposed technical solutions are shot down directly by managers in Sec 4.4.1. In Sec. 4.3.2 we show that some practitioners see fixing bugs as core to their duty and identity as an engineer, which is indeed often an explicit part of their job role, so when they are explicitly told not to *not* fix a bug, this creates what Berti and Simpson call a *double bind* [33], which they explain is a form of disempowerment exercised through *coercive* power.

However, sometimes we see how institutional power affects what practitioners felt able to do and say less directly. We see this in the sense of alienation and psychological toll felt before and after raising concerns in Sec. 4.4.6, perhaps from resisting a systemic face of power Fleming and Spicer call *subjectification*, which "determines an actor's very sense of self, including their emotions and identity" [91], and thus is may be stressful to resist incentives which they have internalized into one's identity as a "good employee". On a similar note, in Sec. 4.5.2, we see how "friendly" bosses and workplace can make it harder to raise concerns, because of a feeling of not wanting to disappoint, which may also function as a form of *subjectification in* organi-

zations, where one is induced into "aligning [one's sense of] self with the organization" [91]. We also have seen in Sec. 4.5.1 how financial and immigration precarity left some unwilling to raise concerns, leaving them to keep their discussions within "boundaries of appropriate ... behavior" [153]. In explaining how institutional control operates, Lawrence and Buchanan [153] do note that employees who are "professionally mobile (based on skills or family connections)" may be better able to resist this control, as we see in this finding.

We also have seen how practitioners seek to resist institutional control, or exercise power of their own. For example, in Sec 4.4.2 we see practitioners seek to negotiate within organizational incentives and persuade others of their concern, as one practitioner sought to set the terms of debate to be about resources, not ethics. Under Fleming and Spicer's framework, this is a form of power they term *manipulation* involving "agenda setting", which "often relies on rhetorical and persuasion skills, and perhaps most importantly, access to key social networks" [91]. Lawrence and Buchanan position this as a manner of exerting agency within institutions called *influence* [153], involving "tactics, including moral suasion, negotiation". In examining how practitioners may be able to "climb the ladder" until their their concerns are taken seriously, we turn to Meyerson's scholarship on "tempered radicals" [182], who Fleming and Spicer also turn to as an example of "counter-subjectifying tactics" which they categorize as a face and site of power as *subjectification against* organizations.

In Sec. 4.4.3, we see various graduations of refusal: from soft refusal enacted by slow walking projects or doing bad work, to overt forms including demanding a reassignment, delivering an ultimatum, to following through with quitting. These forms of agency are hard to locate within the three organizational science perspectives on power we draw on. However, they seem to fit best as an example of an actors' agency within institutions in the face of what Lawrence and Buchanan call "bureaucratic force", when "corporations fire employees", where our participants had accepted the risk of this force nonetheless [153]. Lawrence and Buchanan also recognize softer forms than "direct refusal, but rather an indirect subversion" of company interests which

speak to the examples of soft resistance [276] we also observe.

It is important to note that these frameworks remind us of many faces of power that did not show up in our empirical results, yet exist in wider discourses around ethics in tech. For example, Fleming and Spicer point to other forms of power "against organizations", such as the articulation of new ideologies to challenge industries to challenge the *domination* of hegemonic ideologies. In tech ethics contexts, this may look like critique [108, 142] of the common "Responsible AI" principles which constitute hegemonic "Ethical AI" discourse [137], and articulations of alternative values for tech ethics such as indigenous approaches to data governance [51]. In Sec 4.4.4 we see our participants also engage in "feet voting", proactively planning their individual career choices in a way that align with ethics.If this kind of action was taken on a wider and collective scale, it may constitute what Fleming and Spicer characterize as a site of *coercive* power *against* organizations, pointing to work showing how "social movements mobilize valuable resources to pressure change in firms", and in our case, practitioners' own skilled labor is the valuable resource in question. On this, Berti and Simpson write about how unions may be one way to mitigate organizational disempowerment, by "restoring ... collective capacities to .. voice issues and expose contradictions" [33]. While we do not observe collective action in our data, perhaps due to our attention to non-tech firms (see below), collective action across the technology industry is increasingly prevalent and studied [191].

### 4.6.2 The power to declare an "ethics bug" and dedicate resources to fix it

Smaller ethical concerns, which were often described as "bugs", represent scopes of concern where a technical fix is possible, at least in theory. Ethics interventions such as toolkits [156, 217], checklists [165], principles [137], and education [90] are often designed to help practitioners identify issues, and flag them to others using artifacts such as model cards [184] or datasheets [98]. However, these interventions practically depend on practitioners having the power to dedicate resources, make design changes, or otherwise fix concerns these interventions

may help identify. Without this power, these interventions risk being insufficient at best. At worst, such interventions risk limiting critique to the narrow scope of system design thereby allowing companies to avoid scrutiny of business practices [108], enforceable regulations [196, 259], or fitting into a simple narrative where morally unimaginative engineers are the core problem and training to find ethical issues the solution [286]. Our work shows what happens after practitioners identify concerns without these ethics interventions – and discover severe limits on their power to affect change as they attempt to resolve their concerns.

Similarly to how our interviewee cited academic papers on inclusivity in VR to legitimize her concerns when raising them to her team in Section 4.3.2, other work suggests that fairness checklists may "empower [...] individual advocates" [165], and other tools may enable "uncomfortable design discussions" [120] about gender bias in software design [48]. These tools legitimize ethics concerns, in part by framing them more palatably as improvements to a product (i.e. as fixing "bugs") to improve its chances of success [127, 276], as some of our participants couched their ethics concerns within organizational incentives (see Sec. 4.4.2) and are occasionally successful (e.g. crane software in Sec. 4.3.2).

However, we show that even when less-threatening, narrowly-scoped issues garner agreement that a concern is legitimate, these concerns are often nullified using the usual logics of "customer centricity", as in Section 4.3.2 when inclusive VR was dismissed as something the customer did not demand, or when management or clients dismissed the two technical solutions proposed to remedy concerns as requiring too many resources to implement, as in Section. 4.4.1. In these cases, incentives won out, even for concerns aimed at improving the product rather than critiquing its entirety.

Therefore, fixing "ethics bugs" often relies on practitioners' *power* to persuade others to dedicate resources to fixing them, and this in turn motivates further work to develop tactics of persuasion such as justifying solutions to ethics problems in terms of organizational incentives (see Sec. 4.4.2, see also [276]), and work to quantify and provide outside evidence for rela-

tionship between ethics fixes (i.e. accessibility, see Sec. 4.3.2) and the incentives that decision makers care about, such as product success or user growth. Our work also helps answer calls for "guidance around how to navigate organizational power dynamics" [277] when raising ethical concerns that toolkits help identify, by helping understand the power structures into which ethics interventions must work within, and the limits on the power of those who may apply them. Additionally, practitioners may themselves have power to prioritize among bugs [111], and our work suggests the opportunity to examine where "ethics bugs" lie in their prioritization. Others show that practitioners advocate for ethics less powerfully due to career concerns [165, 216], we show this is inflected by financial and immigration precarity (see Sec. 4.5.1, also [44]) and workplace culture (see Sec. 4.5.2). This suggests future research should investigate other contingencies on practitioner's power to advocate for ethics.

This also has implications for education. A recent survey of undergraduate tech ethics courses found their "overarching goal [...] appears to be to teach students to recognize ethical issues in the world" [90], but fewer than one quarter touch on the systems of power – "capitalism, financial models, marketing, pricing" – within which issues must be addressed. Our study enumerates ethical concerns that practicing software engineers face in Section 4.3.1, which can help ensure in class examples are representative of the concerns that practicing software engineers face at work. However, lest courses help students identify issues but leave them unprepared to advocate for fixes, the factors we enumerate in Section 4.5 can help tech ethics teach students how they may encounter these systems of power in their future careers, alongside learning tactical skills to raise concerns, which we detail in Sections 4.4.1,4.4.2, and 4.4.3.

### 4.6.3 Labor as counterpower to question an industry's *raison d'être*

Practitioners also raised larger concerns which question the *raison d'être* of their organization or industry. Other scholars have critiqued design-stage interventions as insufficient [96], especially when harm is inherent in how systems are used (see Chapter 2, aligned with calls for ethics

work to "move away from prioritizing notions of good design" and towards critique of "what and whose goals are being achieved" [192]. Concerns we collect at this end of the spectrum provide myriad examples of practitioners raising these critiques. Even with the above ways to improve tech ethics interventions, the kinds of ethics concerns addressable using them are likely to remain limited to those aligning with the company's incentives. Therefore, additional research and education is needed to account for ethical concerns which may threaten a company's *raison d'être*.

Our empirical evidence demonstrates that when practitioners develop concerns with their company or industry's business practices, they see few options other than withholding their labor (i.e. resigning and finding a new job, see Sec. 4.4.3). Though this made some feel less culpable in harm, some believed they would be easily replaced and the system still built. Indeed, Palantir CEO Alex Karp said "I've had some of my favorite employees leave" over the company's contract to provide software to US Immigration and Custom's Enforcement that helped separate immigrant children from their families [15, 35], but the contract continued. However, we hear from other participants that resigning can be powerfully disruptive, leading some clients to cancel precarious projects (see Sec. 4.4.3). Future research should explore individuals' power over outcomes: when is a resignation by a concerned engineer successful in halting a software project? This can build on research examining volunteer open source projects, which studies how the departure of crucial "truck factor" developers often puts the project into serious peril [23]. Some of our participants practice "feet voting" by proactively planning their career in alignment with their values (see Sec. 4.4.4), future work can evaluate how commonly and with what priority ethical concerns factor into tech job seekers' priorities, and examine support and information needs for ethically-concerned job seekers.

Practitioners' concerns with their industry's *raison d'être* also has implications for education. Very few of the 115 tech ethics courses surveyed in one study encouraged "students to create their own personal code of ethics" [90]. Given that our work shows that practitioners see their

employment choices as opportunities to exercise agency in accordance with their ethical views, tech ethics courses may consider providing help with career planning as a primary opportunity to align their labor with their values (see Section 4.4.4), in addition to teaching about ethics-focused tech worker rights organizations such as the Tech Workers Coalition [4]. Tech education may also expand to teach skills identified in our study, such as negotiating for ethics using organizational incentives (see 4.4.2), but more powerfully, it can also call attention to strategies for building collective power, including watercooler talk to socialize concerns (see 4.4.6), whistle blowing and legal remedies (see 4.4.5) to discussion of tech worker unions (see 4.4.4).

### 4.6.4 The coherence of a focus on "AI" or "Big Tech" in tech ethics discourse

The power of labor is strongest when acting collectively: as one of our participants recognized, "the work doesn't get done without us" (see 4.4.4). However, only two of our participants raised unions as an avenue to advocate for ethics concerns (see Sec. 4.4.4) despite high-profile efforts to collectively organize over ethics issues at large firms such as Google or Microsoft [146]. Our study shows show that tech ethics research ought to: firstly, broaden to consider tech ethics beyond its contemporary focus on AI; and secondly, broaden beyond studying software engineers at "Big Tech" companies. This larger focus will examine more contingencies in tech worker power and enable a broader coalition by finding issues of common concern, but also shift consideration of ethics towards harm irrespective of implementation, instead of a privileged focus on "AI" concerns.

Firstly, to capture as wide of a scope of ethical concerns as possible, and given divergent conceptions of what "AI" is [145], we did not limit our study to "AI" practitioners, or to concerns related to "Ethical AI". While one analysis concluded that "activism" by "the artificial intelligence (AI) community" was "successful" in part because of "a coherent shared culture" bourne of attending the same conferences, and concluded that "The AI community is acting together

– it is organised" [29], we argue this casts the AI community as a monolith, characterized by its most privileged academic members, and the sources of power and concerns they hold. Despite not deliberately recruiting AI practitioners, most of our interviewees were "building 'smart' machines" in some way (i.e. as per [73]), and some positioned themselves as working on "AI" systems. Despite this, none of our "AI" participants consider themselves "organized" nor talked about themselves as part of a wider shared "AI" culture.

Given this, we argue that a focus on "AI" in tech ethics discourse implies a limited scope of scrutiny, focusing on design-stage interventions [96]. Using "AI" is a design choice, and whereas many of the concerns our participants raise do not depend on whether the system in question uses "AI" or not, especially when concerned with the *raison d'être* of their industry (see Sec. 4.3.2). Therefore, future work on the ethics with software practitioners should avoid limiting recruitment to AI practitioners or framing questions to exclusively AI concerns, as such a limitation may be artificial and limiting in the same way that AI principles may limit scrutiny to system design [108]. The scopes we present in Section 4.3.2 may help conceptualize practitioner concerns, beyond AI. Similarly, "AI Ethics" [40] courses may consider expanding to study tech ethics broadly, as some already do [90].

Secondly, only one of our interviewees currently works for a "Big Tech" company (i.e. [223]), though he did not speak of concerns working there, and only one other spoke of concerns from past experience working in Big Tech. The majority of our interviewees were contingent contractors, working in a variety of B2B companies, or working as software engineers at non-tech companies (see Table 4.1). This is relevant in light of calls to do research beyond "large, internal software development teams" [250], but also given that many ethics issues "are important but arcane and not conducive to media coverage [...] in particular for low-visibility AI companies, including those that do not market to the public but instead sell their AI to governments or other companies." [58]. Major companies invest heavily in certain framings of AI ethics to the point they raise concerns of *capture* of not only AI resources [265] but also AI ethics discourse [281],

and they also are the site of the most high-profile examples of countervailing collective organizing [16, 54, 146], and thus their workers may well be aware of (certain versions of) broader ethics discussions. Therefore, we argue that studies of practitioner ethics challenges, which often focus on "large U.S.-based technology companies" [276] or "major companies" [216] risk assuming a base level of exposure to AI ethics discourse, and thus risk assuming a certain level of generality around what ethics concerns exist.

Given this, we suggests that AI Ethics research may need to broaden to better account for the majority of software practitioners who do not work at "major" companies. For example, we believe that our participants' feelings of being isolated in their ethics concerns and resulting mental health consequences (see Sec. 4.4.6) and attempts to build this community by socializing their concerns (see Sec. 4.4.6) may reflect unique isolation in contrast to in tech-centric companies, where processing concerns with similarly aware colleagues may help [245]. To account for this, and to find ways to build collective power across diverse experiences, future work on software practitioners' ethics concerns ought to deliberately recruit from beyond tech companies, perhaps using existing catalogs of collective actions from a broad variety of tech workers including blue and white collar tech workers [190, 191].

## 4.7 Conclusion

In this chapter, we report on the ethical concerns software engineers identify themselves, without the use of ethics interventions such as fairness checklists [165], codified principles [137], or institutionalized ethics programs [179], which others argue impose a limited scope of ethical scrutiny [108, 142]. Our results show that with an open ended scope, practitioners raise a wide variety of ethical concerns, including those which question the *raison d'être* of their company or industry. We examine the strategies practitioners use to seek to resolve their concerns, and the way in which personal precarity, workplace culture, and organizational incentives limit their power to do so. In our discussion, we highlight the centrality of *power*: our results suggest

that ethics interventions, research, and education must expand from helping practitioners merely *identify* issues to instead helping them build their (collective) *power* to resolve them.

| Gender | Highest Degree | Seniority | Sector | Role | Yrs Coding | Org. Size | Concern(s) |
|---|---|---|---|---|---|---|---|
| Male | MS | Sr. | Government | ML Researcher | 10 | 100-499 | inequality, surveillance |
| Male | HS | Sr. | Government | CTO | 35 | 20-99 | surveillance |
| Male | BS | Sr. | Government | CTO | 20 | 100-499 | legal |
| Male | BS | Mid. | Government | Software Eng. | 8 | 10,000+ | security |
| Male | BA | Jr. | Military | Software Eng. | 20 | 10,000+ | military |
| Male | BS | Jr. | Military | Software Eng. | 6 | 1,000-4,999 | military |
| Male | PhD | Sr. | Edtech | CTO | 37 | <10 | privacy |
| Female | MS | Jr. | Edtech | VR Developer | 14 | 1,000-4,999 | accessibility, inclusivity |
| Male | MS | Sr. | Academia | Researcher | 17 | 500-999 | surveillance |
| Male | BS | Jr. | Academia | Researcher | 8 | 10,000+ | research ethics |
| Male | BS | Jr. | Insurance | Software Consult. | 22 | 100-499 | insurance denial |
| (declined) | MS | Jr. | Fintech | Data Scientist | 18 | 10,000+ | inequality |
| N.B. femme | MS | Mid. | Banking | Data Scientist | 12 | 1,000-4,999 | inequality |
| Male | BS | Sr. | Humanitarian | Software Eng. | 10 | 1,000-4,999 | labor exploitation |
| N.B | BA | Mid. | Health nonprofit | Software Config. | 6 | 10,000+ | life safety |
| N.B | BS | Jr. | Security | Software Eng. | 9 | 10,000+ | privacy, labor |
| Male | HS | Mid. | Construction | Software Eng. | 15 | 10-19 | privacy |
| Male | PhD | Sr. | Mobile dev. | Data Scientist | 25 | 100-499 | privacy |
| Male | BS | Jr. | Networking | Software Eng. | 12 | 500-999 | privacy |
| Male | BS | Jr. | Video software | Software Eng. | 6 | 20-99 | manipulation, misuse |
| Male | HS | Mid. | Agriculture | Software Eng. | 7 | <10 | environment, labor exploitation |

Table 4.1: Interview Participant demographics grouped by sector. To protect anonymity, we do not provide participant numbers nor uniquely identify their continents (spanning Africa, Australia, Europe, with the majority in North America) in this table.

# Chapter 5

# How Workers Discuss AI Ethics: Can a Game Provide a "License to Critique"?

This chapter is based on ongoing work done in collaboration with Nik Martelaro, Laura Dabbish, and James Herbsleb.

## 5.1 Introduction and Related Work

Many companies have Responsible AI guidelines, which often revolve around principles like Fairness, Accountability, and Transparency [137]. However, research shows how Silicon Valley cultural norms, such as technological solutionism and normalization of failure limit those who lead these programs as they seek to enact change [179]. In the meantime, high-profile incidents illustrate these limits. In separate incidents, Meredith Whittaker [63] and Timnit Gebru [181] were ousted from their positions at Google after seeking to organize on issues including military drone tech, and bias and environmental impacts of ever larger language models, respectively. In a recent case of self-removal, "godfather of AI " Geoffrey Hinton left Google in the spring of 2023, "so he can *freely speak out* about the risks of A.I" [180] (emphasis added) and not impact his employer while doing so. While different in many ways, these cases demonstrate that there

are limits to the kind of direct ethical critique acceptable within tech companies.

The prolific adoption of Responsible AI standards in companies [137] may initially seem to legitimize workers as they raise ethical concerns [165]. However, other work shows that they may, in fact, have the side effect of setting specific bounds on what is and is not legitimate to raise as an ethical concern. In their work examining standards encouraging environmental sustainability in companies, Christensen *et al.* show how standards can enact "discursive closure", which they understand in light of Deetz [75] to mean legitimizing certain narrow kinds of employee critique while tacitly ruling others out of scope [57]. They suggest that a "license to critique" must be deliberately created to work against this limiting of discourse, so that such standards can be flexible and enable discussion of concerns they do not specifically enumerate.

Analogous arguments exist for the effect of Responsible AI standards. For example, Keyes *et al.* satirically argue that a system can be Fair, Accountable, and Transparent yet still mulch elderly people into milkshakes, showing how dire harms may be outside of the scope of discursively closed principles [142]. Greene *et al.* show how Responsible AI standards place questions of how to *design* a system as needing the most scrutiny, but "rejecting critiques of business practice" in their focus on system design [108]. In this sense, principles that define responsible AI standards may comprise the *de facto* language of AI ethics, thereby making it harder to articulate concerns outside of this language, both for workers in companies and for activists and policymakers seeking to influence companies' actions.

Others have examined how teams implement AI ethics guidelines (*i.e.*, [137]) into organizational processes [216]. For example, some examine checklists as a way to enable team discussion on designing fair AI systems [165]. This study found that some workers were concerned that advocating for AI fairness issues may impact their career advancement, or lead to them being "labeled a troublemaker", but found that a checklist may be able to "empower individual advocates" to raise issues that are legitimized by the checklist [165]. Related work on UX professionals seeking to steer their company's values has also found using a set of "soft" tactics based on

larger "discourses and logics of the technology industry," allowing these teams to change values while still operating within perceived acceptable bounds of their company [276]. Furthermore, many AI ethics tools or guidelines are often focused on a subset of technical machine learning topics [277]—such examples from CSCW include systems for leveraging crowds to develop ethical constraint specification of AI systems [169] or observing how datasheets for datasets may support ethical thinking [42]. However, given the concerns about the discursive closure effects of standards and sets of principles we outlined above, we question whether enumerated lists of principles, checklists, or other tools rooted directly in technical aspects of machine learning can support workers in raising broad and varying ethical concerns. Additionally, the wider culture in which they are enacted—organizations with certain notions of "efficiency", technological solutionism, and status hierarchies based on technical merit [179]—may limit what concerns workers feel able to raise under these kinds of interventions. Checklists and other artifacts which seek to operationalize AI ethics standards may enact discursive closure, limiting discussion to those issues they enumerate, rather than enabling a broader license to critique. Prior work analyzing AI ethics toolkits find that they often frame the work of AI ethics to be narrow technical work, rarely engaging with wider social issues, or the power dynamics in which this work must take place [277].

To this end, we ask the research question: **RQ1: What factors appear to influence members' "license to critique" when discussing AI ethics with their team?** While many have interviewed AI practitioners individually about ethics issues and processes [122, 165, 257, 258], group dynamics influence how discussion proceeds. We have only identified one study which examines *group* discussions of AI ethics, however, this study was *a priori* scoped to Fairness [164], forclosing discussion of wider concerns as discussed above. Understandably, answering questions about group discussions through direct observation is difficult to study—often, by their nature, AI ethics conversations in companies involve proprietary information and ethical or legal issues that may be highly sensitive. Even in other contexts, such as in activist groups, discussions

may be hard to observe as they are often unplanned, spontaneous, or involve sensitive plans that the group may be unwilling to reveal.

To overcome these challenges, we asked existing teams who have experience discussing AI ethics in their team to discuss AI ethics scenarios in a hypothetical context created for our study (described below). This is done to minimize the risk of revealing sensitive information. We recruited three teams across two companies and one activist group. Inviting real teams allows them to bring their associated shared experiences and context, shared understanding of process, and power dynamics into the study discussions. Including an activist group provides a point of comparison from which to question norms in company contexts. In individual follow-up interviews, we used participants' experience discussing AI ethics in this hypothetical context as a probe to enable them to reflect on differences and similarities between it and AI ethics in their ordinary team context. In short, we seek to learn about how organizational and team norms influence discussion of AI ethics, by using a hypothetical context to (a) enable participants to speak more freely in contrast to sensitive company discussions, and (b) serve as a probe that participants can compare to their past experience.

In designing a hypothetical context to facilitate AI ethics discussions, we also seek to study factors that may help create a license to critique within these discussions. In his book *Domination and the Arts of Resistance*, James Scott drew from his fieldwork to argue that people speak and act differently depending on power differentials between them and their audience, with less powerful subjects using "public transcripts" when in earshot of the powerful, while persistently using "hidden transcripts" when speaking "offstage ... outside the intimidating gaze of power" [228]. Scott emphasizes continuity between these two stages, in particular, that "rumors, gossip, folktales, songs, gestures, jokes" are where people may dissent more freely while "hiding behind anonymity or behind innocuous understandings" [228].

Motivated by Scott's concept of offstage talk [228], we see a connection with games based around speculative futures as a way to provide an "innocuous" context for discussion. We look

toward the literature on speculative futures and the power of speculative games to create more playful contexts which may help resist discursive closure. In their book *Speculative Everything*, Dunne and Rabby articulate how using speculative futures exercises can allow teams to "explor[e] alternative scenarios" to enable them to "be discussed, debated, and used to collectively define a preferable future" [82]. Mankoff *et al.* articulate the value and methods of Futures Studies within human-computer interaction, and in particular the value of "critical reflection" to examine "the relationship between present-day realities and potential futures", mentioning the possibility of fiction and multiplayer games to support this critical reflection [171]. In a technology context specifically, *Project Amelia* used immersive theater to encourage participants to reflect on their privacy behavior within technology [234]. There are also at least two existing examples of games proposed for AI ethics contexts. Ballard *et al.*'s *Judgment Call* was designed around Microsoft's articulated ethical principles to create "space for difficult or uncomfortable conversations" [28]. Martelaro & Ju's *What Could Go Wrong?* is a game where participants combine cards outlining a particular scenario with cards naming a particular user group or exceptional circumstance, and discuss concerns starting from these new combinations. To examine the possibility for these kinds of games to create the "innocuous" understandings that Scott wrote of [228], and to examine how they may work against discursive closure [57], we pose another research question: **RQ2: How do AI ethics discussions unfold while playing a game oriented toward speculative critique?**

We next describe how we observe four existing corporate and activist teams as they play the *What Could Go Wrong?* game, and conduct one-on-one follow-up interviews to compare and contrast their conversations during the game with their perceptions of their past typical discussion of AI ethics. We use the game to provide a point of comparison for participants, to allow them to more easily reflect on their ordinary discussions of AI ethics, and how they may or may not feel license to critique (RQ1). We find that notions of "scope" bound the kinds of concerns that can be raised in AI ethics discussions, and how this is inflected by group power dynamics. We then

look at the specifics of the conversations in the game (RQ2). We find that a game context can broaden conversation, but that games may be unlikely to lead to change directly but may help teammates better understand each other's critical orientation and thus may help form collectives for future action. Our results help AI ethics research better account for team power dynamics, and have implications for research where games are framed as interventions [270].

## 5.2 Methods

We engage three teams from companies and one group of activists to first play the *What Could Go Wrong?* game, and then follow up with one-on-one interviews to probe on differences between conversations had during the game and their past AI ethics discussions.

### 5.2.1 Procedure

Given that we seek to examine how games affect AI ethics discussions rather than build a game ourselves, we choose to conduct the present study using Martelaro *et al.*'s *What Could Go Wrong?* a game in which groups of 4–5 participants discuss a series of AI applications and potential harms. Other games for AI ethics discussion exist [28], but we choose this one because its source materials are readily available, can be easily adapted by adding cards, and includes an online version for remote participants,[1] and because it is modeled on popular party games *Apples to Apples* or *Cards Against Humanity* which may make it more easily understood by participants.

In this game, players first select a random *Prompt* of a particular automated technology (*e.g.*, "autonomous food delivery") and then each chooses one *Response* card which includes particular user groups (*e.g.*, "blind user"), events (*e.g.*, "non-consenter infringement", "random crashes") or exceptional circumstances (*e.g.*, "locusts") to create scenarios in which to discuss the game's eponymous question. A "Card Czar", who does not play a *Response* card leads the discussion by

[1] `https://github.com/nikmart/what-could-go-wrong-ai`

either choosing one *Response* to match with the *Prompt* or by discussing all cards played. The round ends when the group decides to move forward and the Card Czar chooses one card as the winning combination.

Gameplay sessions lasted 1.5 hours, there was no time limit set for each round, with groups completing between three to eight game rounds each.

### 5.2.2 Participants

We recruited 17 participants across four teams, all who had prior history of discussing ethical issues in AI, existing norms of interaction, and a shared organizational context. The first two teams were from a US-based multinational technology company; the first team works on research and engineering for an AI-enabled hardware deployment designed to observe and aid workers in manufacturing environments, and the second provides ethical evaluation and guidance for products and services the company develops. The third team worked at a European-based multinational media streaming company, all of whom conduct research and development work to build algorithmic features and evaluate them for possible ethical issues. The fourth team included members of an activist collective focusing on raising awareness of carceral technology developed and deployed in the city where they are based. Their participation provided a contrasting set of team norms, less influenced by strict hierarchies or tech company practices. Company teams played the game over a video conferencing platform using an online card table simulator[2], and the activist group played in person using a printed deck, reflecting how each of these teams ordinarily meet together. While team C1-T2 played with their manager (which we discuss in Section 5.3.3), no other company teams did, but we note significant diversity in team seniority (which we discuss in Section 5.3.3) as can be seen in Demographic details for each group provided in Table 5.1.

[2]www.playingcards.io

### 5.2.3 Data collection

Each team played the game in a session lasting 1.5 hrs. The conversations for each team were audio recorded with consent and IRB approval. Except for one participant who declined, all participants participated in a one-on-one semi-structured [264] recorded follow-up interview (lasting an average of 34 minutes, but as long as 49 minutes or as short as 22 minutes), asking them to reflect on in-game experiences and conversations which arose, but with special focus on contrasts between in-game discussions and past AI-ethics discussions within their team or organizational context, and on the extent to which they did or did not feel comfortable raising anything, or disagreeing, during in-game or prior AI ethics discussions (see Appendix XX for the interview guide). For remote sessions, researchers turned off their cameras unless answering questions to minimize any effect their presence may have had on teams' discussion.

### 5.2.4 Data analysis

We analyzed data under an interpretive epistemological paradigm [160] using deductive thematic analysis [59] . We began by open coding [242] a selection of six transcripts of two play sessions and follow-up interviews selected for diversity of role, company, team, and past experience, annotating portions relevant to our research questions, while collating coded portions and associated themes into an analysis document. After reoccurring themes emerged in this document, a tentative code book [59] comprised of five initial codes was constructed, and applied to all transcripts. In approximately weekly meetings between authors, new codes arose to capture new themes of interest, in which case the coding frame was updated and data re-coded in an iterative process. Two categories–those addressing our research questions specifically–had a large amount of divergent data, so coded portions were printed out and sorted into finer grained cohesive categories using an open qualitative card sort [284].

## 5.3 RQ1: What factors influence members' "license to critique" when discussing AI ethics with their team?

We find that organizational norms push against critique, in particular through a notion of "scope", and that the power and critical orientation of those in the room affects whether participants feel able to bring up ethical issues. Here, we rely primarily on data collected during one-on-one follow-up interviews, while their experiences during the game primarily serving as a probe to encourage reflection on ordinary discussions of ethics.

### 5.3.1 Organizational norms push against ethical critique

Across company participants, a feeling of organizational norms, often implicitly understood, modulated whether they feel able to bring up ethical issues during their typical conversations around AI in their work. For example, one participant reflecting on past deliberation around AI ethics noted an expectation that there should be a *"significant enough concern" (J)*C1-T2 before they would *"have a conversation about it" (J)*C1-T2, suggesting that raising issues should be saved for only the most dire cases.

Other participants relayed how it felt difficult to raise ethical critique about new technologies, in the face of wider company excitement. She said that *"just saying no [¡about a product idea, redacted¿] just makes everybody frustrated [...] striking that balance is something I'm still learning how to do properly. And so it takes some work [and] conscious effort." (L)*C2-T1. She gave the example that during "large forum" company meetings, when someone is presenting new technology that she might have ethical issues with, there is often *"a lot of enthusiasm going in, [which] I think make[s] it hard to kind of speak out. [...] you've got like all these like, emojis like 'thumbs up', 'loving it', and then like the chat is blowing up with people saying how amazing [the tech] is." (D)*C2-T1.

Another participant discussed how she was self conscious that negativity might go against

company norms and forward progress for their team, relaying that she had been told that she is *"too negative at work. And don't focus enough on the positives [...] It's difficult to pull that back" (E)*C1-T1, but that she was recently *"trying to be a 'team player', and really trying to hold back when I disagree." (E)*C1-T1 After being told by a close confidant that *"you complain a lot [...] you shouldn't do that." (E)*C1-T1 she has *"been trying to ramp back on disagreements and save it for when I feel most passionately." (E)*C1-T1

### 5.3.2 "Scope", its contestations, and its effects

In our results, we saw the notion of *scope* invoked to bound or express what a corporate team believes it can or will take action on, and thus where discussion is focused. It appeared as a softer way to limit bounds of discussion—saying a statement is "out of scope" is not a judgment that it was incorrect or imaginary—but that it was beyond what a group believes organizational norms or incentives would permit them to discuss or take action on. We found that scope is often enforced through reference to time pressures, a push to solve particular problems often through technical means, and through role divisions leading to compartmentalization of ethical questions.

#### Scope: broadness or narrowness of critique

Various notions of "scope" surfaced as key attributes defining what participants consider acceptable to raise in work discussions about AI ethics. For example, in contrast to game discussions, one participant said, *"a lot of the discussions I've been having recently have been much more narrow" (L)*C2-T1, often focused on specific aspects or features of a product. One participant suggested that her work discussions about AI ethics don't have the *"sense of freedom to go off and think about very unlikely harms that could happen and discuss those further" (K)*C1-T2. She went on to say how she would like to integrate some of her outside "passions" into AI ethics discussions at work, for example "dystopian" themes from *"watching shows like Black Mirror and reading all of these, you know, sci-fi dystopian stories" (K)*C1-T2, but *"those are things*

*that I probably wouldn't share in an Ethical Impact Assessment review" (K)*C1-T2, because it wouldn't be "applicable."

Participants perceived that certain kinds of harms or remedies, such as climate harms or those requiring systemic change "diffuse" across society can often be seen as out of scope and hard to discuss. For example, one participant relayed how someone brought up in a large team setting how *"LLMs have like a major climate impact. [...] And you try to bring these things up. [...] it's kind of falling on deaf ears that are unwilling [to hear this,] they're just kind of like, 'nope, we're being told use LLMs, [so] we're using it."' (D)*C2-T1 Another suggested that *"AI ethics is that it often falls short of trying to work towards systemic change." (R)*C2-T1 One participant spoke about how she feels most AI ethics conversations don't talk about more "diffuse" cultural effects, elaborating *"Like what does it mean for people to use technology kind of pervasively in a specific way? [...] we don't, as often, I think, talk about [this]" (F)*AC.

However, some ethical issues were seen as so potentially damaging to product success, that they cease to be ethical questions, instead expanding in importance to become "product questions", as one participant relayed: *"big enough problems that [...] it's beyond the ethical questions. It's also just a product [question]" (L)*C2-T1. This suggests that something being an "ethical question" may be its own kind of limiting or minimizing scope.

**Time pressures and questions of relevance**

Participants reflected concerns in follow-up interviews about how time pressure scopes discussion only to ethical issues seen as most "relevant". Multiple participants reflected on time pressures within their teams, showing how this served to continually foreclose discussion of less direct yet pressing–in the eyes of participants–concerns. Others suggested how engineers on their broader team may perceive AI Ethics conversations as lacking relevance to their work. One said, that some engineers thoughtfully think through ethical issues, some engineers push back with questions of relevance: *"sometimes the pushback of 'oh, that's an edge case, that's never*

*going to happen."' (D)*C2-T1. Others suggested that there might not *"be too much enthusiasm" (J)*C1-T2, because it would be perceived as *"a big chunk of time without a great ROI [return on investment]" (J)*C1-T2, or that some might not *"see how it would [...] be applicable to the work" (RB)*C1-T1 they do.

**Push to "solve": discursive closure by being scoped to "fix" a particular system**

In a similar sense to what Christensen *et al.* call "closure by design" in sustainability standards [57], participants wrote about how in past AI ethics discussions, a "goal orientation" affected, and to an extent, limited, the kinds of conversations which arose. In some ways, this makes sense: participants in companies often were tasked with evaluating existing or proposed systems for ethical concerns, and then suggesting mitigations—largely technical changes to the system —to (partially) resolve those concerns. In other ways, techno-solutionist or techno-chauvinist thinking can entrench existing inequities, and obscure other ways of thinking or conceiving of problems or solutions [69]. Along these lines, some participants reflected on how the "problem-solution" script could preclude discussion of wider changes, such as systemic change, and does not fit into their prescribed role.

Some talked about how AI Ethics conversations are often prompted by a specific product or problem: *"maybe we're talking about large language models [and the] impacts of generation on, like, artists [...] So it's a little bit more targeted" (L)*C2-T1, comparing this to in-game discussion which *"felt more generative [because] there wasn't as much of a goal" (L)*C2-T1. A participant also noticed how conversations often *"jump towards like, what's, what's a good technical solution?" (L)*C2-T1, continuing to say that most past conversations were *"solution-oriented, [like] if we're looking for a mitigation, what's the best or most practical way that [...] we can do that [...]you're trying to get at a smaller solution space" (L)*C2-T1. In this way, narrowing the scope of discussion was viewed as a process of "solving the problem".

Participants even spoke of how internal ethics tooling and processes lead to discursive clo-

sure: *"the tools internally, they're a bit more guided [saying] 'if you're interested in building a system or model, here are a bunch of questions that we want you to answer' [...] tend to be a lot more directed" (L)*C2-T1, for example asking narrower questions about whether a system she might work on uses protected demographic information. She mentioned, therefore, *"They're a little bit less broad in terms of [...] societal impacts off [of our] platform. [...] the focus feels a little bit narrower." (L)*C2-T1 A participant from a different company spoke of this too, suggesting that conversations in-game were more "creative" than when dealing with "reality" when *"having all the details, like we do [when we do] an Ethical Impact Assessment" (J)*C1-T2

Participants suggested a variety of possible reasons for this rush to discuss technical solutions. One suggested that in a *"tech company" (D)*C2-T1, with an *"engineering mindset" (D)*C2-T1 open-ended introspection is *"not always the vibe" (D)*C2-T1. One participant reasoned that this may be due to *"what is possible to change" (R)*C2-T1 within an individual worker or team's power, but also due to a cultural mindset biasing towards being *"able to measure particular harms, the things that are not measurable, end up not being as easy to solve for." (R)*C2-T1

**Role divisions leading to compartmentalization of ethics**

Some participants discussed how role divisions affect who is expected to care about, and handle, AI ethics questions. For example, one said that this wasn't his job, saying these issues were handled by a specialist committee, and a member of his team who *"target or address those topics on [our] projects[...]we have people that do that" (G)*C1-T1. However, many others were concerned about this apparent "compartmentalization" of ethics: *"the compartmentalization of what we do with any individual horizontal capability, I think this is a huge problem with respect to ethical uses of AI" (RB)*C1-T1. Participants spoke about countering this compartmentalization, saying they *"need we need more diverse thoughts here" (RM)*C1-T1 and seek to *"strengthen the bonds among some of the product, ML product practitioners, and me [in her ethics-focused role]" (R)*C2-T1.

### 5.3.3 The power and critical orientations of those in the "room"

**If managers (or others with power) are in the room**

More than just demonstrating a general awareness of who is in the room and how that affects what is safe to share, participants appeared acutely aware and concerned with how bosses and managers shape the conversation. One noted that compared to other conversations, the game was a space where: *"your boss isn't here" (E)*C1-T1 nor was the session being recorded by her employer. Therefore, *"you're free to talk about things that you think are weird or risky." (E)*C1-T1 Another participant said that in the play session *"the things that I discuss here, it's not going to impact my paycheck next month. So it's more comfortable" (RM)*C1-T1. One participant commented on the *"surveillance technology that's on everyone's [company] laptops" (R)*C2-T1, and also on *"worker exploitation" (R)*C2-T1 during the play session, but noted she wouldn't feel comfortable *"bringing [this] up when it's not just around peers [and if] we had managers in the room [who are] on the company side." (R)*C2-T1

Some participants suggested that disavowal of their critique was sometimes justified by managers using an ostensibly altruistic rationale, saying *"'I'm trying to make your life easier,' we don't need to do this." (D)*C2-T1. Others reported their *"manager, and like, my skip level[ managers]" (K)*C2-T1 encouraging her to prioritize work that *"they felt would be more impactful in a product" (D)*C2-T1. However, this did not always appear to be the case. In one play session including a manager, their subordinates said *"I don't feel like there's really censoring that goes on or filtering if you will." (J)*C1-T2, and another said that given their past experience working in formal ethics team, *"we're all peers [...] I knew I could share freely in front of this group of people." (K)*C1-T2 Thus, while hierarchies may impact what some team members feel they can discuss, specific team norms may help to support all team members in speaking more freely.

**Awareness of teammates' Critical Orientation**

Participants displayed awareness of who was in the room, and what their ethical views and critical orientation, and reflected on how that may have influenced how they expressed themselves, especially when raising certain kinds of critique. One participant said this directly: *"who is in the room can change the tenor of a conversation and can change the tenor of how you deliver critiques or hold back critiques" (E)*C1-T1 She went on to say she'd frequently discuss concerns like "privacy" and "fairness" with all members of her group, but discuss concerns like AI displacing human labor with only a subset of them: *"there's some people in the group, who are, whether by virtue of their discipline or their interests, are more attuned to [...] discussing things like labor" (E)*C1-T1

Others suggested that they habitually discuss AI ethics topics among their team, but that *"the dynamics [...] probably would [...] be different if it was, like, any of us, with people from other teams" (RB)*C1-T1, because they wouldn't have a *"shared baseline" (RB)*C1-T1. Another participant stated *"it would have taken me longer to sort of establish sort of as an internal feeling that people were sort of engaged in a discussion in good faith." (L)*C2-T1. Even those on AI ethics teams may not feel completely aligned with their direct team when shared views of the critical questions around AI are *"not so sharply in focus" (D)*C2-T1, as they were on their past teams.

Some were worried about particular consequences arising from raising critique around unfamiliar people, including that this could *"impact [...] who I [can] collaborate with [...] some people can be really sensitive." (RM)*C1-T1 One participant suggested that webs of social and collaboration networks are opaque in companies, leaving her unwilling to critique other researcher's projects. Others reflected on raising specific topics during the game due to the (dis)comfort with their team. One company participant said she felt comfortable raising topics like worker exploitation because she knew *"it's a group of [...] like-minded people." (R-Act)* In a different example, an activist participant felt pressure to stay "on topic" and raise critical points, because

her fellow players were such a *"critical group of people. I might [otherwise] have been goofier in playing a card game. [...] more like trying to[...] just like fuck around" (A)*AC.

### Personal attributes and status hierarchies

Participants recognized and discussed how gender, seniority, and level of technical experience affected the status one might have in a particular room, and thus the license with which they felt able to raise critique, or affected group dynamics in such a way that made raising critiques feel more or less possible.

We observed that age and gender affected perceptions of who is able to speak up. For example, one participant, lamented that his younger colleague *" was not really talking up [speaking up]. He was not grabbing time" (RM)*C1-T1 because *"he is really young. And [...] he joined very recently" (RM)*C1-T1, In contrast, another very senior participant joked *"you know me, I talk about anything. Maybe when I was younger, I might have been more cautious." (RB)*C1-T1. Another participant whose play session included female-identified people suggested *"there's a different way these conversations happen in all-female groups than when there are other genders present [...] men take up space in particular ways" (R)*C2-T1 Such comments suggest how age, seniority, and gender may affect perceptions of who can or should speak up.

Participants also spoke about their roles within engineering organizations and their backgrounds. One female-identified participant with a non-engineering background stated: *"if an engineer seems to be saying something that I think is wrong, I don't know, he's an engineer, and he's been here 20 years, maybe I'm wrong" (E)*C1-T1, suggesting how seniority, "engineering" expertise, and perhaps gender, may impact who is perceived as "wrong" in company contexts. Participants in the activist group also noted if they should engage in critique while not having an engineering background. A female-identified member of the activist group suggested her lack of computer science expertise may be a shortcoming, saying *"if I was in the room with people who were developing AI, I might feel uncomfortable just because I don't have the same depth of*

*knowledge on the topic as they do." (R-act)* Another participant felt they might not qualify to participate in this study *"I'm not one of the requirements in [study criteria] was to be like in an engineering field. So I was like, I'm not that." ()*AC

Some participants, many of whom had graduate educations, also reflected on the status of those with academic backgrounds and how this can quell critique in AI ethics discussions: *"whenever there's some very senior professor speaking [...] people don't speak out against them [...] people tend to agree" (RM)*C1-T1.

One participant noted that their personality and the amount he speaks may lead others to agree too quickly, "overpowering others" *(C1-T1)*.

## 5.4 RQ2: How do AI ethics discussions unfold while playing a game oriented toward speculative critique?

Relying primarily on observations and recordings of the game session, here we examine how the game was able to expand scope, how participants remixed rules, and used the game context as an opportunity to learn about teammate's past experiences and critical orientation.

### 5.4.1 Expanding scope

#### Randomness as scope expander

Participants found that the randomness provided by the cards and game rules could be a valuable way to expand their conversations and critiques, especially beyond what they might normally discuss. One participant whose work focuses on the ethical challenges of content recommendation reflected that *"it was cool to [be] outside of the [content] recommendation space for a second [...] Because you can get into a rut [and] having like a new [example] helps you see some of the gaps [...] in your own thinking [that you're ] habituated to" (R)*C2-T1. Participants from other sessions corroborated this, one stating *"the format of giving responses [cards] ends*

*up, like, forcing you to make connections that maybe you wouldn't have thought about before." (L in session)*C1-T2. Another mentioned that a format where *"there's not really a correct answer" (L)*C2-T1 is one she hadn't considered before, but noted that it *"got a lot of us thinking in different directions" (L)*C2-T1.

Participants also suggested that subjective interpretations of the same card served to expand the scope of discussion. One said this directly: *"people came up with things I didn't expect, despite looking at the same card" (L)*C1-T2 A member of the team from the activist group suggested that "randomization" helped expand scope which was useful for a different reason, as *"usually when I have conversations like this they're about a very specific real thing, right there, like either something's happened in public in the news, or something that someone's working on something [concerning]. They're not necessarily like speculative." (A)*AC, thereby helping drive conversations about speculative possible futures that need not be reactive to any particular news event. Other participants also expanded the scope of discussion by integrating parts of their own lived experiences during in-game discussions, integrating discussion from outside of the particular set of cards at hand. One participant referenced how technology had changed street culture in her native India by putting the "juice man" on the corner out of business as people moved to app-based delivery services, another relayed about protests against visual noise wrought by advertising on metro trains in her native Saint Petersburg, a third relayed how her native Berkeley was "awash in Kiwibots" with their "pixel heart" eyes, and fourth spoke about their experiences on a team where a robot had physically harmed someone. As one member of the activist group reflected: *"all of us had such a different frame of reference" (F)*AC, and people appeared to feel space to speak from this frame of reference throughout game play.

**Thinking beyond the product**

Participants also found that the game led them to consider scope beyond the product and towards second or third-order harms. One participant stated *"[we] were thinking like a couple of steps*

*ahead [to] society at large, whereas [discussions] in practice tend to be about a more narrow, so like [...] how is this product, harming users in ways that are measurable and quantifiable? " (R)*C2-T1 Similar sentiments were echoed by members of the activist group. For example, one participant reflected that in her life, she usually discusses concerns about AI within the context of a specific AI system "actively happening" in the present or recent past, and appreciated the opportunity to talk about future possibilities: *"[our discussions] were more theoretical in the sense that we weren't talking so much about a specific form of AI [...] So it was interesting to kind of talk about it in a more intangible way. Although it's always about, you know, kind of predicting the future in an intangible way." (R-Act)* Another member of the activist group echoed this separately in their own follow up interview, suggesting that their in-game discussion spoke to *"these cultural intangibles that really got to, I think, the deeper root of some of our concerns." (F)*AC

**Hypothetical situations as an "innocuous" context for discussion**

Several participants brought up how a hypothetical context in the game, which may provide an "innocuous" context, as suggested by James Scott [228], to raise critique that may otherwise be too socially costly to raise. Given randomly drawn prompts and dealt response cards provided, one participant suggested if the "structure of the game" "pushes" one to *"bring up things that you [otherwise] wouldn't feel comfortable bringing up" (L)*C1-T2 then *"in that context, it probably does make it easier" (L)*C1-T2. Reflecting on other hypothetical interventions she's participated in before, this participant also reflected that this makes it easier for people not just to raise critique they may have but be nervous to raise, but accept and themselves raise critique of things *similar* to their own work while being less "defensive": *"once we did it in a hypothetical sense, people were looking at this and going 'oh, okay, well, yeah, it's not about whether we intended for something to go wrong [...] things can really go wrong!"' (L)*C1-T2, saying that this allowed people to "disconnect" from the frame of *"'something that you're doing is incorrect', or 'there's*

*something unethical about your work"' (L)*C1-T2, that when conversation is is *"taken away from the project that we were doing [...] everyone was very free" (RM)*C1-T1.

Others corroborated this, suggesting that raising concerns *"hypothetically in a game [...] is really nice because it's just it's a lower barrier [...] versus talking about a specific project which [...] is going to be much more serious and have potential real-life implications right as you bring up different concerns" (Session)*C1-T2 One said the point is to "make assumptions": *"I saw this one is, you know, if you I mean, I'm making a lot of assumptions here. But that I think that's maybe the point of some of these discussions. " (Session)*C1-T2 Another participant remarked how she appreciated the opportunity to *"take ourselves out of, you know, okay, 'this is a real product that we have to provide actionable guidance and feedback on' to [instead] 'okay, let's just have our, you know, brain flowing to think about all the possible what ifs, what could go wrong with this scenario."' (K)*C1-T2

**Hard to transfer from hypothetical context to real world action**

However, we found that this hypothetical context may make it difficult for in game discussions to transfer to real world action. Importantly, this raises questions about whether discussions rooted in in-game hypotheticals can spur real-world action. Reflecting on past AI Ethics trainings based on hypotheticals, one participant found them effective, but noted: *"The minute you start talking about their projects, you see a very different behavior. [...] They're very concerned about these projects showing up in a negative light. And that that being I mean, people start to become more defensive. They don't expand into all the things that can go wrong." (L)*C1-T2

Revisiting a quote from the first part of this section, one participant said: *"hypothetically in a game [...] is really nice because it's just it's a lower barrier [...] versus talking about a specific project which just by nature is going to be much more serious and have potential real life implications right as you bring up different concerns" (Session)*C1-T2. Among these "real-life implications" may be the idea of real world action, such as through existing compliance

processes. However, one recommended against this: *"I could see possible resistance is if it's seen as a checklist activity. So if it's perhaps tied into like a compliance process, and like, you must do this before your product goes out the door, then there could be some resistance there" (C1-T2)* This suggests that it may be difficult to fuse the hypothetical context created with the game with an integration with requirements for mandated changes to actual products.

A participant in a different group echoed this, reflecting on the good conversation from her groups play session, and wishing there would be a way to translate this into action: *"My complaint [...] with team based [...] conversations... like when two people talk [...] the whole is greater than the sum of the parts. [...] But translating that thing that is made into something that is captured and can be operationalized [...] has been a consistent issue." (E)*C1-T1 This was also apparent in the words some participants used to refer to the session, one calling it a *"non-work space [...] almost like a team building exercise" (R)*C2-T1, which appeared to set the expectation that this is not the context from which immediate or actionable changes to product or process in work contexts would arise.

### 5.4.2 Learning about teammates

#### Vulnerability and space to socialize

Participants spoke about how the game context made certain conversations possible that they felt otherwise unable to have. In one instance the Response card "Random Crashes" prompted a participant to share that he had previously worked on a robot which had killed its operator. This was the first time he had shared this with his teammates despite them working on physical systems and frequently discussing safety and ethics concerns. In follow-up interviews, his colleagues reflected on this: *"[he] shared with us that he was working in this factory, where actually a robot did a "random crash" and [...] killed somebody. [...] It was impact, like it was, it was really shocking. Like 'Oh, wow' like he was part of it. Like he was there." (G)*C1-T1 Another participant suggested that with an *"all-audio [meeting] culture" (E)*C1-T1 that *"It was nice to*

*get that kind of space where we could more like, really talk about things we were seeing or things that we were thinking about [that were] not necessarily constrained by our work" (E)*C1-T1.

**Learning others' critical orientation and finding allies**

Apart from sharing sensitive past experiences, some spoke about how the vulnerability prompted by the game created a unique opportunity to learn about a teammate's critical orientation. By *critical orientation*, we refer both to their values and perspective on ethical issues in technology, but also their willingness to critique project's goals or company incentives, when this may be in conflict with the former.

For example, one reflected how in a *"non-work space, but [where we were] still be able to have conversations that are adjacent to what we're doing [...] helped me see that we're more or less all on the same page [and] who my allies are in this in this fight." (R)*C2-T1 On a similar point, one participant in another company stated how she had previously discussed more critically oriented topics, such as labor displacement, with only some of her coworkers, but had *"probably self-selected out of discussing certain things with [other] folks due to [their] backgrounds [or] presuming that they're not interested" (E)*C1-T1. However, reflecting on the game session, she relayed how she appreciated hearing from teammates *"with whom conversations can be very tight and narrow, to hear them pontificating a little more, engaging in [an] imaginative exercise. [...] I've only ever heard them talk about dialogue prompts [so] it can be easy to assume [that they] don't think the same way that I do [...] theory of mind can be difficult to achieve." (E)*C1-T1 In her view, this game provided an opportunity for her to learn how more members of her team felt about more critically-oriented topics.

In a follow-up interview, one of the members of the activist group discovered that she had similar concerns to a member of the group she had only met briefly, and after playing the game noted: *"I really like their perspective [,...]some things that they said [...] made me want to talk to them further." (A)*AC

We note that exposing one's critical orientation, especially around those you do not know well, may be a risky endeavor: one risks being labeled a trouble maker [10] or concerns of career repercussions [165]. We return to this further in Section 5.5.1, below.

## 5.5 Discussion

Our results show that in game and in ordinary work, discussants seek to understand and display sensitivity to both the differentiated power and critical orientation of their discussion partners, which they use to modulate AI ethics issues they choose to raise and how they present them. When one's boss is in the room, or colleagues of unknown critical orientation, people may be less willing to raise critique. This echoes the work of James Scott, demonstrating how people employ "public transcripts" when those with power over them are present, but use more frank offstage talk when speaking to teammates they trust [228]. Additionally, a variety of factors affect people's perception of their own status, such as their seniority in the team or their proximity to engineering knowledge, in turn also affecting their willingness to raise critique.

The most straightforward implication of this finding is that those designing future AI ethics interventions intended to be used in a group discussion context must attend to the differentiated power relations of discussants. Explicit attention to this may include exercises for a group discussion to begin by reflecting on these, reflexively discussing these as a group, or even simply an enumeration of what kind of power relationships (*i.e.*,, boss/subordinate, as well as those related to age, seniority and gender) to look out for. Naming these things will not level them, but doing so is already more attentive to power dynamics than many existing AI ethics interventions. Our work joins a great deal of other work [38, 96, 138, 154] which makes clear that research on AI Ethics must be more attentive to the differentiated power and relationships of power in those that may use, request, or engage in proposed AI ethics interventions or when discussing AI ethics issues (see also Chapter 4). If future empirical work examining how those discussing AI ethics in any context do not attend to power in their analysis, such work risks missing major determinants

of any apparent agreements or disagreements that may arise.

### 5.5.1 Hypothetical game context may not lead to change directly, but it may help find critically-aligned allies

Our work casts doubt on whether the innocuous context, such as those created by games, may enable discussions that transfer to changes in a team's real-world context. While Scott's suggestion that "rumors, gossip, folktales, songs, gestures, jokes" are the places where people may demonstrate dissent more freely by "hiding behind anonymity or behind innocuous understandings" [228] and suggest that game-based AI Ethics interventions may expand the scope of what is sayable "on-stage" and create such contexts to make dissent more safe, our results tell a more complicated story. As we detail in Section 5.4.1, participants spoke about how they felt able to speak freely specifically because the context was hypothetical: not connected to a particular project and not tied to a particular "compliance process," as one participant said, which may demand politically difficult or time-consuming changes to one's product. This gap between the hypothetical context and "real-life implications," as one participant put it, is both a powerful attribute of the intervention—it is the feature that made specific conversations possible that were not before—but also a powerfully limiting factor of the intervention, in that this gap was seen as being maintained by *not* implying any change outside of the hypothetical context.

Our findings therefore cast doubt on whether discussions or agreements during in-game contexts may transfer back to action in business contexts, where this would imply real work, real shifts in direction, or sign-off from higher-ups. Given this, our findings render a critique of the framings of AI ethics games that purport to spur real-world action. For example, when proposing the game we study, Martelaro *et al.* claim that a "little lightheartedness can promote more productive conversation about otherwise negative topics" [175], and Ballard *et al.* found that "having a serious conversation about ethics and technology in the context of a game creates space for difficult or uncomfortable conversations. Within this conversation, the use of design

fiction to create discursive space [...] deflects blame or charges of irresponsibility in actual settings with actual harms" [28]. While our results suggest that discursive space may indeed have been created, it is still unclear and unknown how such conversation may lead to averting actual harm from real work.

This being said, our results suggest a more subtle but perhaps enduring mechanism of action for games to shift organizational realities—finding allies by developing an understanding of their critical orientation through gameplay. Our results suggest that the hypothetical context fosters vulnerability, such as through sharing sensitive past experiences working on AI systems that had caused deadly harm. Such stories help reveal parts of team members' critical orientation and allow others to learn about their critical orientation. When these personal understandings and relationships transfer to real-world contexts, this may help form coalitions to address real-world "actual harms." Scott emphasizes continuity between the two "stages" [228] he proposes, and relationships that form "off-stage" appear to be the conduit towards enabling "on-stage" solidarity. Another participant discussed how they had felt comfortable discussing possible labor displacement implications of AI systems they were themselves building, conversations which they had not previously had with certain members of their group, presuming some were not interested in such topics. Reflecting that in ordinary work contexts, "theory of mind can be difficult to achieve", she relayed how she appreciated learning more about her teammates on topics they didn't usually discuss through this "imaginative exercise". Another participant in a different group reflected on how this game helped her learn "who my allies are in this fight". While strengthened social ties, or one or two more allies may seem small, "if subordinates are entirely atomized, of course, there is no lens through which a critical, collective account" can emerge [228], and we join Scott to suggest that collective accounts are where solidarity begins.

While formal AI ethics activities such as checklists may be able to "empower *individual* advocates" (emphasis added) by legitimizing a particular issue [165] contained on a narrowly scoped checklist, intersubjectivity developed through gameplay may enable individuals to better

know one another, who their "allies are" in the words of one participant, from which a broader collective to raise critique may be fashioned, less constrained by the discursive limits of any particular standard [57]. Chapter 4 demonstrated the severely limited ability of employees to raise concerns beyond a very narrow scope, and instead suggest that future "ethics interventions, research, and education must expand from helping practitioners merely identify issues to instead helping them build their (collective) power to resolve them, and our results here suggest that "innocuous" contexts (*i.e.*, [228]) created by games may provide space for collective power to begin to form. In organizational psychology, this concept is termed "cross-understanding", defined as "the extent to which team members understand the other members' mental models" [132], and while the literature on this construct often focuses on the impact of cross-understanding for "product quality" and to avoid cases where group members may make "proposals concerning the group's processes and product features" that other members would find "technically, politically, or otherwise unacceptable" [124], parallels may be drawn beyond quality and features, to questions of product ethics.

Feminist theory helps illuminate the distinction between hypothetical or "innocuous" [228] contexts enabling real-world changes directly versus enabling stronger ties, which then become a powerful basis from which action may then arise. Donna Haraway argued that we ought not to suppose that there is a "view from above, from nowhere", and thus that trying to suppose a context that can create one, is both unlikely to succeed and may be harmful [115]. Extending Haraways argument, Lucy Suchman argues that responsibly developing technology must be a "boundary-crossing activity, taking place through the deliberate creation of situations that allow for the meeting of different partial knowledges" [247]. Rather than theorizing games as separate safe spaces from which to speak from nowhere, we suggest that games may be opportunity to deliberately allow different partial knowledges to meet, learn what they have in common, and enable "collective knowledge of the specific locations of our respective visions", from which durable coalitions and collectivities for action may arise. Drawing on both Haraway and Such-

man, Chapter 3 showed how social ties, responsibilities, and concerns, developed outside of engineers' assigned duties are the basis for AI ethics work that does get done, innocuous contexts created by games may enable non-work contexts these ties to form and strengthen.

This has implications for game design research, especially if intending to intervene in group dynamics for prosocial ends, particularly in contexts like workplaces with built-in power hierarchies. Such work may consider framing their game as an opportunity for relationship and coalition building more so than a context where direct changes to real practice will arise. This may include examining the effect of any such intervention and examining contingencies on the durability and outcomes from any resulting relationships formed, over a longer time span.

### 5.5.2 "Out of scope" as a rhetorical device to softly dismiss critique

Our results show how notions of "scope"–received notions of team believes it can or will take action on and thus they bound discussion–are constructed and maintained, how this limits what is considered acceptable in AI ethics conversations, and ways that participants sought to say things outside of these bounds (see Sections 4.3.2 and 5.4.1). Some of our participants discussed how **individuals compartmentalize ethics**, in ways that limit what they perceive as in scope during AI ethics discussions, with participants from companies suggesting that out-of-work experiences or passions are not in scope for AI ethics discussions. In contrast, those in the community activist group did not feel this way. Additionally, some of our participants discussed how labor is divided in ways that leave ethics to be the primary remit of one team member, leaving others feeling that ethics issues are beyond their own "scope." Some of our participants also segment critique between projects, in order to avoid perceived career consequences when working with different team members. In these ways, the wholeness of any individual's perspective is itself compartmentalized, leading to a narrowed scope of discussion when teams meet. This elaborates what was argued in Chapter 3, demonstrating how ethics is modularized between team members and within individuals as they choose to bring only *fragments* of their own partial perspective to

these discussions.

Secondly, others have discussed how tech industry logics such as technological solutionsism affect ethics initiatives [179], our results illustrate how notions of **efficiency become a scope limiter**, casting a subjective assessment of priorities in the more objective language of "scope". Participants' direct references to their calendar and scarcity of time, and to less direct notions of relevance or framing some harms as "unlikely", lead to a situation where only possible harms perceived as most relevant, or most likely, are seen as most in scope and thus most legitimate to raise for discussion.

Thirdly, given these time pressures, teams report how discussion is most often scoped towards **that which feels actionable, often technical changes,** in line with how technosolutionism operates to limit discussion to immediately actionable fixes. This is evident in how participants describe their workflow as being presented with particular systems they are supposed to evaluate for ethical issues in company contexts, and propose "mitigations" to solve said issues, and how this already can make it harder to discuss how the systems they're evaluating relate or many require "systemic changes" questions this current process does not present as part of their agency to discuss.

Finally, our results suggest how **scope can be tested or expanded**, in how affordances from game-like interventions such as randomness may give social permission to do this 5.4.1, and how people may employ rhetorical moves to frame ethics questions as larger scoped product questions.

We theorize scope, in many cases, as a softer way to dismiss critique. This functioned as an instance of problem closure [125], whereby "rhetorical process through which relevant social groups perceive their problems with an artifact to be solved or closed", but in a way that softly dismissed those who may wish to keep it open or believe it to be unsolved. Casting an issue as "out of scope" merely says that it is beyond the team's remit or practical ability to act on, without forcing one to contend directly with the issue raised. By avoiding denying the validity

of the issue outright, and thus avoiding dismissing the validity of a colleague's sincerely raised ethical concern. This is a particular instantiation of *jurisdictional stasis* – a concept in rhetorical analysis with its roots in classical Greek, but with more modern adaptations to questions of "moral decision making or ... practical concerns" [253]. Instead of arguing that an issue is false, arguments based on jurisdictional stasis question the "jurisdictional appropriateness of the issue" [253], that is, whether an issue is within the jurisdiction or scope of a particular group or team.

While not always described as such, many scholars have theorized how standards, principles, and toolkits seeking to guide organizational behavior toward "pro-social" are discursively closed. In their analysis of environmental sustainability standards, Chistensen *et al.* [57] demonstrate how they risk discursive closure: *closure by the past*, where responses to future problems are limited by standards developed for past concerns; *closure by design*, where overly-prescriptive standards leave no freedom for adaption and become a putative "seal of approval"; and *closure by routinization*, where standards are solidified into organizational processes in ways that are difficult to change. In an AI ethics context specifically, Greene *et al.* analyze AI ethics statements of principles, examining how they "legitimate (and delegitimize) certain practices", finding in part that by focusing on how to design AI systems rather than the business practices they enable, they frame "business practices [as] being discursively 'off the table"', implying that "'better building' is the only ethical path forward"' [108]. Keyes *et al.* satirically demonstrate how narrowly scoped "Fair, Accountable, Transparent" design principles scope scrutiny to system design, warning against "treatment of ethics as a series of heuristic checkboxes that can be resolved technically" and thereby avoiding engagement with "wider societal issues" [142].

This suggests that designers of future AI ethics interventions ought to see the risk of discursive closure and deploy particular ways to reduce this risk. Our results suggest particular design affordances that may help do this. While the designers of the 2020 Microsoft AI Fairness checklist recognized the risk of discursive closure in that they included disclaimers in the checklist's

extensive preamble like "Undertaking the items in this checklist will not guarantee fairness. The items are intended to prompt discussion and reflection" [166], our results suggest that more than written disclaimers or warnings are needed to avoid discursive closure. More extensive changes to an intervention's form and including deliberate design affordances, such as randomization, are needed to resist this closure.

## 5.6 Conclusion

Past work has sought to design AI ethics interventions–such as checklists [165] or toolkits [36]–to help practitioners design more ethical AI systems. However, other work demonstrates how these interventions [277] and the principles they're based on [108] may serve to instead limit critique to those addressed within the intervention, while rendering broader concerns illegitimate.

In this paper, we draw on work examining how standards in other contexts enact discursive closure [57] and work on how power relations affect whether and how critique is raised [228] to examine how teams discuss AI ethics issues. We recruit three corporate teams, and one activist team, each with prior context, to play a game designed to trigger broad discussion around AI ethics, and firstly use this as a point of contrast to trigger reflection on their teams' past discussions, examining factors which may affect their "license to critique" in AI ethics discussion. We then report on how particular affordances of this game may influence discussion, paying particular attention to hypothetical games as a viable mechanism for real world change. We discuss how power dynamics in a group and notions of "scope" affect whether people may be willing to raise critique in AI ethics discussions, and our finding that games may not be able to lead to direct change, but may be more likely to allow members to find critically-aligned allies for future action.

| Resides | Citizen | Gender | Current Role | Yrs in org | Highest Degree, Field |
|---|---|---|---|---|---|
| Company 1, Team 1 *(C1-T1)* — Remote Session | | | | | |
| USA | USA | Woman | Research Scientist | <1 yr | PhD, Social Sciences |
| USA | USA | Man | Research Scientist | >25 yrs | PhD, Social Sciences |
| Germany | India | Man | Research Scientist | 1-5 yrs | PhD, Computer Sciences |
| Mexico | Mexico | Man | Research Engineer | 1-5 yrs | MS, Computer Sciences |
| USA | India | Man | Research Scientist | 1-5 yrs | PhD, Computer Sciences |
| Company 1, Team 2 *(C1-T2)* — Remote Session | | | | | |
| USA | USA | Woman | Program Manager | 5-15 yrs | MS, Humanities |
| USA | USA | Woman | Program Manager | >25 yrs | MBA |
| USA | USA | Woman | Director | 15-25 yrs | MS, Computer Sciences |
| Company 2, Team 1 *(C2-T1)* — Remote Session | | | | | |
| USA | USA | N.B. Femme | Research Scientist | <1 yr | PhD, Social Sciences |
| USA | USA | Woman | Research Scientist | 1-5 yrs | PhD, Computer Sciences |
| USA | USA | Woman | Research Scientist | 1-5 yrs | MA, Humanities, Social Sci. |
| USA | USA | Woman | Research Engineer | 1-5 yrs | MS, Computer Sciences |
| Activist Collective *(AC)* — In-person session | | | | | |
| USA | India | Woman | Masters Student | 1-5 yrs | BA, Arts |
| USA | USA | Woman | PhD Student | 1-5 yrs | BA, Humanities, Arts |
| USA | Russia | Woman | PhD Student | 1-5 yrs | MA, Social Sciences |
| USA | USA | Woman | Designer | <1 yr | MS, Arts |

Table 5.1: Participant demographics. To protect anonymity and reduce re-identification risk, some descriptors here have been generalized, and we provide group rather than individual identifiers alongside quotes.

# Chapter 6

# Discussion and Implications

My dissertation has found that organizational norms and boundaries limit software creators' sense of responsibility and agency over the downstream impact of what they create. I have also examined the possibilities and shortfalls of play as a way to resist these limits, and found that while offstage games are unlikely to lead directly to onstage change, they may provide an opportunity to forge stronger relations for onstage action. In this section, I discuss the implications of this dissertation for tech research, interventions, education, and policy.

## 6.1 Research Implications

### 6.1.1 Relations and Scope

My research demonstrates how dislocations between units–individuals, teams, organizations–shape what ethics work gets done, and what is ruled out of scope. The *supply chain* metaphor–briefly introduced in Chapter 2 and developed in Chapter 3–zooms out from a particular context to demonstrate these dislocations, but also allows us to situate social and technical units in the context of their social and technical relations, thus rendering these relations visible for strengthening or scrutiny.

The supply chain metaphor has implications for how tech ethics researchers choose what to focus on. The link between one's chosen unit of analysis is already understood to be a fundamentally important aspect of research study design, for example, in case study research [280]. However, my findings may help tech ethics researchers better understand how these other kinds of scope they may adopt in their study–implicitly or explicitly–may determine the kind of results they are likely to find. While the boundaries between technical systems, teams, or organizations may seem like reasonable boundaries to scope research studies, my findings in Chapters 3 and 4 show that things at these boundaries often affect what ethics work does and does not get done. Narrowly-scoped studies on tech ethics–for example, focusing on the challenges of a particular team developing a particular system–may inadvertently represent issues found in terms of the work the team focuses on, and miss relations at organizational boundaries which my research shows can be core to to ethics work.

My research also demonstrates particular kinds of relations that future tech ethics research should be sensitive to. These include customer demands, public perceptions or pressure, individual's friendships and kinships, how developers (do not) relate to users or data subjects, as I examine in Chapter 3; power relations between managers and workers, precarity, and organizational incentives, as I examine in Chapters 4 and 5; and software licenses as I examine in Chapter 2. My work shows that these relations shape how ethics concerns are conceived of, legitimized, and acted upon, contextual factors which may be missed with a too narrowly-scoped unit of analysis, such as that of individual decision makers or a specific team.

Relations should not be seen as messy outside influences, but instead tech ethics research should consider relations at unit boundaries in scope for study, or even a primary object of study. This may look like dissecting the "anatomy" of the wider system, including its labor, data, resource, and funding relationships following Crawford And Joler [68]. This may also look like using Dumit's exposition on Haraway's method of inquisitively "imploding" a particular artifact to examine labor, epistemological, material, political, economic, historical and other dimensions

of "*connectedness*" [81]. Finally, research may be inspired by approaches drawing on Suchman's argument for responsible technology development as the "entry into a network of working relations" and the partial knowledges within this network.

Practicality may demand a preordained, particular conception of what is in or out of scope for study, as may be the case under some research paradigms [280] or organizational pressures. However, attempting to position and acknowledge the unit's technical, social, and organizational relations, and acknowledging these in the writing, however briefly, will help position these relations as visible and contestable in future study and critique. For example, if tech ethics research adopts a methodology that does not include scrutinizing the system's relationship to business practices, hopefully this can be acknowledged in how it is written and discussed, in order to avoid rendering such critique implicitly out of scope and invisible.

Change in the real world is often a stated goal of tech ethics research, and some argue it is rarely achieved [96], but understanding relations which preclude or enable change will help make desired change more likely. Past work argues that tech ethics research orients towards "the environments in which technologists research, design, and develop digital objects and systems", but without understanding the (lack of) relations in separate deployment contexts [96]. This suggests parts of tech ethics research may suffer from its own rhetorical closure: prematurely "perceiv[ing] their problems with an artifact to be solved or closed" [125]. Is an ethics issue solved when a technical solution is shown to exist, or when that solution is shown to be effective within the relations of a real organization? I explore this further, below.

My dissertation motivates greater tech ethics research attention to open source development contexts. A long line of research seeks to understand practitioners' challenges doing tech ethics work in company contexts[76, 122, 165]. However, my research in Chapter 2 shows that assumptions in corporate contexts do not hold in open source contexts committed to the sharing and release of code. For example, companies *can* avoid releasing a system until ethics issues have been remedied (in theory, even while many do not). This suggests that tech ethics re-

search should broaden to include open source development contexts. Importantly, however, the boundary between corporate and open source contexts may not be so clear. the open source community I studied in Chapter 2 is independent from corporate influence, but many open source AI projects have corporate sponsors, stewardship or employee contributors. Future research should investigate the intersections of corporate and open source norms in corporate-backed open source projects: how do the corporate norms and incentives we study interact with open source norms we study, and how does this inflect AI ethics possibilities and limits?

Future research should examine the ethical impact of different "gradations" between fully open and closed [236]. Among these are "ethical source licences"[1] or "behavioral use licences" [64], which seek to proscribe uses of the licensed code or model that the licensor deems unethical. There are open questions about whether these licenses are enforceable in practice, concerns about the headaches from multiple licenses with different ethical commitments, as well as resistance from those seeking to preserve ideologically pure notions of openness [237]. However, my research in Chapter 2 shows that licenses and the communities which adopt them can both set norms around acceptable use. Future research should examine software communities' processes of adopting them, as well as how effective they are both in a literal legal sense and in setting norms down the supply chain.

Finally, as I suggest towards the end of Chapter 4, the term "*AI* ethics" may itself imply a another kind of narrow scope that future research should interrogate, insofar as the deployment of AI is one design choice among many which may achieve similar goals. Using this term in research may have its own effect in limiting efforts to design side interventions, a focus already implicitly embedded and excavated in scrutiny of AI ethics guidelines [108].

[1] See: `https://ethicalsource.dev`

### 6.1.2 Practitioner Responsibility, Agency, and Power

My work helps answer calls that center questions of power in tech ethics research [38, 96, 196], instead of reducing ethics to technocratic notions of building better technology [108, 142]. Among such calls, my work helps answer a call to understand the power of "who gets to make decisions?" [155]: my work traces the contours of how people building software systems understand their responsibility for and agency over the impact of what they create, both between organizations (primarily in Chapter 3), within organizations (primarily in Chapter 4), and within teams discussing ethics (primarily in Chapter 5). While some research argues that "The AI community is acting together – it is organised" [29], and paints this community as broadly powerful, my work shows that this community is not so homogeneous. In many cases, my work shows how it is rarely practitioners themselves who have the authority to make consequential decisions on questions of ethics, even while ethics work is delegated to them.

My research shows how limited visibility between organizational boundaries can make "doing ethics" difficult. For example, in Chapter 3, a lack of visibility into and authority over downstream uses made it difficult for those practitioners I study to understand or feel responsible for downstream harms. Yet, when software engineers did develop ethical concerns of their own in Chapter 4, attention to these concerns were often overruled by countervailing incentives, such as being responsive to market or customer demands. In Chapter 2, we see how broader discourses of Technological Solutionism and Technological Inevitability limit what practitioners feel able to affect, in this case arising from norms underwritten by open source licenses.

Past work has examined practitioner's work practices as they engage in tech ethics work [76, 122, 165, 201], and others have written about how broad silicon valley logics hamper ethics initiatives [14, 179]. My work illustrates continuity between the micro-politics of developers (not) doing ethics work, and the larger forces these developers may not recognize explicitly that play a large role in shaping ethical outcomes. Future research seeking to understand barriers as practitioners do tech ethics work should interpret these barriers within the ideologies under

which they often must operate, and conversely, broader examinations of contradictions in tech discourses around ethics should examine how these contradictions are viewed and managed by those called to do the work. To do otherwise would be to locate accountability incorrectly: on one hand absolving practitioners who have some agency in how technical systems are built for the choices they are able to make, on the other, being blind to the fact that they are "operating within a limited sphere of knowing and acting" [247] that leave other options feeling unthinkable.

Within organizations, future tech ethics research focused on practitioners should examine how they seek respond to larger scoped concerns such as those that challenge their organization's *raison d'etre* that I identify in Chapter 4 , given that my findings show these are often hardest for workers to raise and seek to resolve. Ethics concerns cannot be tactically expressed in terms of improving project success [127, pg 131] or other organizational incentives, when the wider goals of the project or company itself is of ethical concern, as I demonstrate in Chapter 4. In a world where tech ethics is often the onus of individual workers [14], tech ethics research should continue to incorporate perspectives from collective action and labor studies [93, 224, 245], and examine ways of building worker power to advocate for ethics.

On an additional note centering labor, my research in chapter 4 shows that some technologists seek to make career changes in line with their personal sense of ethics, and see this action as a primary opportunity to exert agency over the ethical impact of their work. However, we find that many are only able to understand the ethical impact of a job once they are doing it. Future research could examine how ethics fits into technologists' job seeking priorities, and examine their information needs to support these decisions, including possible labor market transparency mechanisms so that job seekers can anticipate in advance whether a given job aligns with their own ethics.

Within teams discussing ethics, future research should be sensitive to discursive closure– the termination of reflection and debate [57] –following my work in Chapter 5. This chapter showed how organizational norms, power dynamics, and discursive frames brought into ethics

conversations may affect what is and isn't raised and scrutinized. Future tech ethics research examining practitioners' work practices should be sensitive to the discursive norms at work in these practices–how they function to leave some changes feeling possible while others unable to be raised in conversation, and how practitioners may seek to resist or re-enforce this closure. In particular: given mine and others' findings [57, 91] that whether "the boss is in the room" determines what some practitioners will raise in discussion–future research should be sensitive to managerial hierarchies (manager/ subordinate) and other power dynamics, and how "agenda setting" power is enacted, which may be investigated by adopting a sensitivity to onstage and offstage talk [228]. This may involve looking past direct utterances or straightforward interviews, to what is joked about rather than said directly, or what is said in group contexts versus what is left to one-on-one side channels.

The salience of gender–and power dynamics thereof–surfaces through my work, even while this is not the focus of my analysis here (even as it is in some of my other work [95, 266]). All participants developing the deepfake tool identified as men even as the majority of harms from deepfakes accrue to women (as some did acknowledge, see Chapter 2). All participants in my examination of how AI supply chains dislocate ethical responsibility identified as men (see Chapter 3), except one woman, who described how ethics questionnaires were cast as secretarial work, and always relegated to her. When I surveyed 115 people who had ethical concerns with what they had been asked to create (see Chapter 4), ten responses came from women and six from non-binary or gender nonconforming participants, one of whom recounted how she was dismissed when she sought to advocate for more inclusive VR. Finally, when I conducted four game sessions with 16 participants (see Chapter 5), eleven identified as women and one as non-binary. This is notable given that three of these four teams were those specifically formed either to think about ethics in their technology, or to advocate against carceral technology, and in the latter case, many did so from explicitly feminist perspectives. In this study, we also observed how age, gender and technical background influenced who was afforded a "license to critique,"

noting that older, male, and technical people appeared to recognize themselves and be seen by others as being more able to be contrarian. There are undoubtedly many gendered forces at play here: the most basic being the under-representation and marginalization of women in technical fields. Joining this, there is the notion that within the technical companies I focus on, it is technical expertise that is privileged above social science or other expertise more often held by women. And in these companies, I show how ethics is often itself compartmentalized and modularized, seen as work to be assigned to those with particular expertise, rather than everyone's responsibility (see Chapters 3 and 5). In this way, in a world where "ethics work" is secretarial, it may be seen as an instance of "non-promotable work" [26].

Software engineering research may consider contending with the negative ethical side effects of modularity, namely that its technical practice reifies a cultural practice of disavowal, under which thinking about ethical harms not encapsulated within the module is made difficult, as I argue in Chapter 3. This may include radical imaginings of alternative conceptions of programming languages than those based around encapsulation and separation of concerns, and more immediate questions of how to modularize ethical concerns (*i.e.*, explicitly delineating division of ethics labor) or expose interfaces that allow ethical questions to be more easily asked and answered between modules.

## 6.2 For Interventions and Practice

In this section, I discuss implications of my work for practitioners, and those seeking to design tech ethics interventions to influence practice. I note that in some cases, the very framing of an "intervention", as a discrete or packaged action towards change, may undersell the scope of what may be needed–given how deeply entrenched modularity is within software engineering, or profit incentives within firms, for example.

My work shows how practitioners' understandings of their responsibility and agency limit what ethics work is done. Therefore, those designing tech interventions (*e.g.*, checklists [3,

165]s) should first seek to understand the agency and responsibility of those they hope will use their intervention, lest they design an intervention that works in theory but which practitioners feel unable or unmotivated to use in practice. For example, many of the practitioners I report on in Chapter 3 worked in companies with tools and processes designed to enable Responsible AI: but divisions of labor and authority, notions of scale, and other factors meant these interventions were often seen as meaningless. My research shows that practitioners sense of agency and responsibility is informed by their position on the supply chain, decision making power, and proximity to a particular user. By accounting for target users' agency and responsibility, this will help designers ensure the interventions they propose do not only work in the lab [96], but are able to more effectively transfer to changes in practice. Indeed, interventions like checklist or guidelines seeking to work under divisions of labor and uncertain responsibility should seek to delineate ethics labor, making it more specifically clear who is supposed to do which task, in order to localize interventions to different roles within technology companies, and help ease coordination challenges across roles [76]. However, if interventions do seek to delineate labor, there is a risk that this will reify discursive closure in how they are applied, by rendering only very narrow kinds of input legitimate from each role. Following my examination of discursive closure [57] in how tech practitioners discuss ethics in Chapter 5, I suggest ways discursive closure may be avoided in ethics interventions, including the use of informal and hypothetical contexts, and allowing people to speak from their own personal experiences. However, even writing what is or is not in scope for a given intervention (*i.e.*, technical considerations only? customers? business models?) will help, by rendering this visible and more easily contested, rather than this being presented as natural [186]

Some practitioners disavow any responsibility, as I show in Chapters 2 and 3, often underpinned by commitments to permissive software licenses or technical practices of software modularity. As I argued in Chapter 3, modularity functions not only as a technical practice but as an epistemic culture, which often serves to bracket off harms not addressable within the module

practitioners may be working on. Challenging the technical practice of modularity may be a practically difficult exercise (as I discuss above), but future tech ethics interventions may seek to challenge the negative cultural side effects of this deeply entrenched practice, by seeking to combat ethical disavowal which accompanies it. In a company context, this may look like intervening to make connections in supply chains more explicit, possibly using means I show in Chapter 3, including highlighting the power of contractual obligations, marketing, relationships with family and friends, or journalism in intervention design. In an open source context, this may include further development of open source licenses, manifestos, and contributor guidelines, along with other ways of countering the norm of 'for use by anyone for anything' [238]. Platforms hosting open source code could recognize the power they have to set norms or platform policy, and choose to intervene and make demands that projects take steps to enforce policies against harmful or illegal uses, as I show happens already in Chapter 2, at least in an ad-hoc manner.

Tech ethics interventions sometimes assume developers *can* know who will use their technology, and control how it is used. In Chapter 2, I introduce this as "Assumptions of Downstream Control": tech ethics interventions sometimes assume a single organization has control over the development process, or whether and how interventions are employed. For example, Google's Responsible AI tutorial for TensorFlow lists "Who am I building this for? [...] How are they going to use it?" as primary questions to ask before starting to build something [78]. However, my work in Chapters 2 and 3 show how these are often questions that developers are unable to answer. Interventions must be designed to operate under weaker Assumptions of Downstream Control, which often only hold in large, vertically integrated technology companies which control a large majority of their tech stack, and often still only for products deployed to users directly. Obliquely recognizing this, a different document from Google outlining their broader approach to Responsible AI recognizes how questions of downstream control and use are not often answerable, instead asking "how closely the solution is related to or adaptable to a harmful use" and "whether we are providing general-purpose tools, integrating tools for customers, or developing

custom solutions" [208]. Future tech ethics interventions should ideally find ways to spell out the the answers to these questions at the outset, but if downstream control cannot be assumed, help practitioners better think through ways their work may enable or be adapted for harmful uses.

Interventions may seek to help practitioners communicate their concerns, forge relations, or build collective power. As I show in Chapter 4, fixing small ethics bugs often boils down to negotiating with managers to dedicate resources to do so. Past work designing tech ethics interventions seeks to help legitimize this work [165]. Future tech ethics interventions could focus on helping practitioners communicate their concerns, better understand the incentives of those who have power to deploy resources to remedy them, and establish processes or bodies to protect employees as they raise concerns.

However, as discussed also in Chapter 4 larger ethics problems often question the fundamental incentives of a practitioner's organization or industry, and in a world where ethics work is too often left to individual practitioners [277], tools to build their collective power would prove useful here. Therefore, future tech ethics interventions could focus on helping practitioners build relations across the fractures in supply chains that I demonstrate in Chapter 3. In the small scale, this may look like informal networks, outside of company structures, to help practitioners meet more of the "many hands" [194] which have a part in working on related technology. One way of building informal networks outside of company structures may take the form of games, as I explore in Chapter 5, and this work provides a possibly more nuanced and realistic mechanism for action: building relations between partial knowledges, that may form a basis for future actions, rather than assuming agreements in-game will directly transfer to real world action. This may also look like interventions designed to help practitioners engage in "soft resistance", which I demonstrate in Chapters 3 and 4, and as found in past work [276]: slow walking projects, doing bad work, and other methods of quietly subverting company incentives. In the large, this may look like interventions designed to help practitioners unionize their workplace and engage in wider collective action [4].

Despite warnings on the follies of technosolutionistic approaches to tech ethics that I and others make (Chapter 2, also [108, 142, 277]), there can be a partial role for technological interventions to support the creation of ethical technology. In cases where downstream control is needed but cannot otherwise be achieved, technological solutions such as encryption or monitoring [134] may help prevent or detect proscribed uses, respectively. My findings in Chapters 3 and 4 show that practitioners often do not know how their code is used, but often develop ethical concerns about this use once they do. To help discover this use, open source platforms or others may seek to trace the usage of code throughout repositories, to reassemble dislocated software supply chains. This may help practitioners understand how the fruits of their labor are used, and make decisions on what they work on in the future accordingly. This would require much engineering resources to build–but surely no more than did the Amazon Echo I opened my thesis by discussing.

## 6.3 For Education

My research in Chapters 3 and 4 demonstrate real world ethics concerns that practitioners face "in the wild", which are sometimes more banal than those which garner media attention [58]. This may help tailor education towards the kinds of ethics issues and barriers to resolving them which students may face in their future career, and may support students in developing a sense of their own personal codes of ethics (as some courses already do [90]) that is informed by specific ethics issues my research shows that practitioners have faced in the past. Also, my research shows strategies (Chapter 4) and relations (Chapter 3) that practitioners seek to use to resolve their ethical concerns, and while some tech ethics courses do include some training on this–for example, focusing on communication and argumentation skills [90]–my research may help broaden the range of strategies taught to students to raise ethical concerns to managers, and see them resolved.

However, my work shows that a strongly communicated or argued ethical concern is rarely

enough. Ethics education should consider teaching about ways of advocating for ethics that do not depend on persuasion but instead teach students to build counterpower to resist organizational incentives (recognizing that this places a lot of onus on individuals, see Policy interventions enumerated below). My research shows how a significant source of power engineers have is in their labor (Chapter 4), in the absence of being able to otherwise resolve their concerns, participants refuse jobs or find ones which align with their ethical views. Similarly, ethics education could therefore consider including career counseling, to help students find jobs that align with their ethical views or concerns. As some may already, more courses should consider teaching about tech worker collectives and unions, some of which advocate for ethical concerns in addition to more traditional labor issues such as benefits or pay, such as the right to "raise concerns about products, initiatives, features, or their intended use that is, in their considered view, unethical" [4]. Courses should consider including content on more subversive tactics of soft resistance–slow walking unethical projects, for example–as my work in Chapter 3 and others [276] show how these strategies can be effective, if partial.

Tech ethics education should consider framing tech ethics as a supply chain problem, by including content beyond the ways to build "more ethical" discrete technological systems. According to a recent review of syllabi, issues of "fairness, bias, and profiling" within algorithmic and AI systems were a predominant focus in tech ethics courses [90]. This is good, but as my findings in Chapters 2, 3 and 4 show, incomplete. Courses should help students learn to see how systems are situated within a wider supply chain of modules and organizations, each having different intended uses, representing creators' partial knowledges, and imbued with different kinds of "ethical debt" [89]. This will help students understand tech ethics as a wider process where technology has politics [272], and that organizations, relations between them, and their incentives and downstream customers merit scrutiny.

Similarly, tech ethics education should educate students about the power dynamics which may await them in their future employment, as informed by my research. The same survey of

tech ethics courses referenced above [90] found that the majority of courses tend to be designed to help students "recognize ethical issues in the world", with fewer than a quarter including content on "capitalism, financial models" or other systems of power which await them in technology organizations in their future careers. My research can help teach future practitioners to understand how these systems of power may affect their future practice of tech ethics.

Finally, my work shows that engineers may use arguments of technological solutionism disavow ethical responsibility (Chapter 2). Even if this education is designed for future software engineers, for whom technical approaches to solving ethical issues are most likely to be within the authority of their job role, future engineers should be taught to recognize the limits of these arguments, as some educational approaches already attempt to do [21].

## 6.4 For Policy

A key insight of my work is that policy makers should reject self-regulatory approaches, such as those based on AI ethics guidelines that companies often develop themselves [137], because they rely on individual practitioners having authority within their organization to operationalize and implement them, which in Chapters 3 and 4, I show is often not the case. These guidelines [137] are also often narrowly scoped to questoins of design rather than use [108, 142], but my work provides evidence that practitioners' ethical concerns exceed this scope in ways that they do not have the power to address. This motivates the need for regulation to account for these larger scoped concerns about the business uses that new technology enables.

My work in Chapter 4 shows that given the lack of agency to act on ethics within their job role, software engineers instead seek to quit jobs they find unethical and find ones that align better with their ethical views. However, in the past, technology firms have colluded to not hire employees from their competitors, and other actions which the US Department of Justice said "disrupted the normal price-setting mechanisms that apply in the labor setting" [2]. Tech companies have also engaged in union busting [74], foreclosing another mechanism which may help

practitioners resolve their ethical concerns [4]. My research in Chapter 4 also shows that some practitioners on employment-based visas cannot quit jobs even when they develop ethical concerns with what they're asked to do, because of this immigration precarity. Therefore, regulatory and enforcement agencies should examine ways of tracking and better protecting tech workers' right to bargain collectively and raise legitimate workplace concerns without fear of reprisal.

My work motivates a supply chain view of regulating technology companies. OpenAI recently disavowed ethics issues in how Kenyan data labelers identify traumatic content in their datasets, by saying that they had subcontracted this out to another firm and thus these working conditions were not their responsibility [5]. Policymakers seeking to regulate AI should hold companies accountable for conducting due diligence for ethics issues in their software supply chain, analogously to how existing laws seek to hold suppliers of physical goods responsible for supply chain ethics issues like child and slave labor. This regulation may require companies to verify that the datasets were collected with consent, or whether upstream crowdworker data labelers were properly paid. This approach may also look down the supply chain, perhaps taking a "product safety" framing, such as requiring that companies verify that models companies seek sell or use in a specific domain are indeed fit for that purpose (*e.g.*, in medicine or legal contexts), especially as recent examples show that not doing so can cause harm [278]. Other approaches may include demanding supply chain transparency, for example demanding companies reveal which libraries, models, or datasets they use, to permit scrutiny and audits by external third parties. In particularly high stakes (such as recidivism prediction tools) or regulated contexts (such in fair housing advertising), regulation should demand that companies release the models and source code, to aid in audits which policy could also require.

Past research into ethics concerns and how practitioners respond to them tend to focus on marquee US-based technology companies [122, 165]. Some of this scrutiny is warranted, as these companies tend to be economically powerful and also drive a fair amount of AI ethics discourse through their issuance of AI ethics guidelines [137]. However, past policy research has suggested

that many ethics issues "are important but arcane and not conducive to media coverage", and noting this may be true "in particular for low-visibility AI companies, including those that do not market to the public but instead sell their AI to governments or other companies." [58]. Across all of my interviews which are the basis of the work presented here, very few (on the order of three) were with participants working in marquee tech companies (so called FAANG companies, *i.e.*, Google, Facebook, Microsoft, or similar). My research demonstrates a wider variety of ethical concerns that software engineers working outside of technology companies may face, and which policymakers should attend to. Software engineers and tech ethics concerns exist outside of FAANG, but may not get as much scrutiny due to their lower profile [58].

My work also motivates expanding policy attention beyond company contexts, where some existing regulatory levers exist, to open source software, where due to its decentralization and lack of central corporate governance structure, traditional approaches to regulation are difficult. This may include scrutinizing content policies for platforms which host open source projects designed for the creation of illegal content, recognizing the power of platforms as I lay out in Chapter 2. However, given corporate control over many open source AI projects–and corporate lobbying for an "open source exception" to recent proposed AI regulations [102]–regulators should not paint open source with a broad brush, instead being attentive to corporate power and interests masquerading as free software, which has traditionally opposed it [62].

# Bibliography


[1] Coalition Against Predictive Policing in Pittsburgh. https://capp-pgh.com/. (document)

[2] Us department of justice v adobe et al., antitrust, 2010. URL `https://www.justice.gov/atr/case-document/file/483451/download`. 6.4

[3] Google will not renew pentagon contract that upset employees. *DrivenData*, 2019. Accessed: 2021-07-7. 2.4.3, 6.2

[4] A tech workers' bill of rights. *Tech Workers Coalition*, 2020. Accessed: 2021-010-1. 2.4.1, 4.6.3, 6.2, 6.3, 6.4

[5] Exclusive: The $2 Per Hour Workers Who Made ChatGPT Safer. https://time.com/6247678/openai-chatgpt-kenya-workers/, January 2023. 1.2, 6.4

[6] Christophe Abrassart and Marc-Antoine Dilhac, 2018. 2.1

[7] Christophe Abrassart, Yoshua Bengio, G Chicoisine, N De Marcellis, M Warin, Gambs Dilhac, et al. Montreal declaration for responsible development of artificial intelligence. 2018. 2.1

[8] Adam Tunnard. CMU custodial staff ratify new contract with Aramark. *The Tartan*, February 2020. (document)

[9] Philip Agre. Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI. In *Bridging the Great Divide: Social Science, Technical Systems, and Cooperative Work*. 1997. URL `https://pages.gseis.ucla.edu/faculty/agre/`

critical.html. 1.2, 3.2, 3.4.3

[10] Sara Ahmed. *Complaint!* Duke University Press, 2021. 5.4.2

[11] Henry Ajder, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen. The state of deepfakes: Landscape, threats, and impact. *Amsterdam: Deeptrace*, 2019. 2.2.1, 2.2.2

[12] Bilal Al Sabbagh and Stewart Kowalski. A socio-technical framework for threat modeling a software supply chain. *IEEE Security & Privacy*, 13(4):30–39, 2015. 2.4.3

[13] Shaosong Ou Alexander Hars. Working for free? motivations for participating in open-source projects. *International journal of electronic commerce*, 6(3):25–39, 2002. 2.3.2

[14] Sanna J. Ali, Angèle Christin, Andrew Smart, and Riitta Katila. Walking the Walk of AI Ethics: Organizational Challenges and the Individualization of Risk among Ethics Entrepreneurs. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 217–226, June 2023. doi: 10.1145/3593013.3593990. 1.1, 6.1.2

[15] Mike Allen. Palantir's CEO said he's suffered because of his contract with ICE. https://www.axios.com/2020/05/26/palantir-ceo-ice-immigration, May 2020. 4.6.3

[16] Anat Alon-Beck. Times they are a-changin': When tech employees revolt! *Md. L. Rev.*, 80:120, 2020. 4.6.4

[17] Alexander Amini, Ava P Soleimany, Wilko Schwarting, Sangeeta N Bhatia, and Daniela Rus. Uncovering and mitigating algorithmic bias through learned latent structure. In *2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 289–295, 2019. 2.4.2

[18] Lameck Mbangula Amugongo, Nicola J. Bidwell, and Caitlin C. Corrigan. Invigorating Ubuntu Ethics in AI for healthcare: Enabling equitable care. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 583–592, Chicago IL USA, June 2023. ACM. ISBN 9798400701924. doi: 10.1145/3593013.3594024. 1.3

[19] Mike Ananny. Toward an ethics of algorithms: Convening, observation, probability, and timeliness. *Science, Technology, & Human Values*, 41(1):93–117, 2016. 2.2.1

[20] Mike Ananny and Kate Crawford. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society*, 20 (3):973–989, 2018. 2.4.2

[21] Aditya Anupam. Playing the belly of the beast: Games for learning strategic thinking in tech ethics. *International Journal of Role-Playing*, (13):31–39, 2023. 6.3

[22] Hannah Arendt. Collective responsibility. In *Amor mundi*, pages 43–50. Springer, 1987. 2.4.2

[23] Guilherme Avelino, Leonardo Passos, Andre Hora, and Marco Tulio Valente. A novel approach for estimating truck factors. In *2016 IEEE 24th International Conference on Program Comprehension (ICPC)*, pages 1–10. IEEE, 2016. 4.6.3

[24] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018. 2.2.1

[25] Ayhan Aytes. Return of the crowds: Mechanical turk and neoliberal states of exception. In *Digital labor*, pages 87–105. Routledge, 2012. 1.2

[26] Linda Babcock, Maria P Recalde, and Lise Vesterlund. Why women volunteer for tasks that don't lead to promotions. *Harvard Business Review*, 16, 2018. 6.1.2

[27] Carliss Young Baldwin, Kim B Clark, Kim B Clark, et al. *Design rules: The power of modularity*, volume 1. MIT press, 2000. 3.1

[28] Stephanie Ballard, Karen M. Chappell, and Kristen Kennedy. Judgment Call the Game: Using Value Sensitive Design and Design Fiction to Surface Ethical Concerns Related to Technology. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, pages 421–433, San Diego CA USA, June 2019. ACM. ISBN 978-1-4503-5850-7. doi: 10.1145/3322276.3323697. URL `https://dl.acm.org/doi/10.1145/3322276.3323697`. 5.1, 5.2.1, 5.5.1

[29] Haydn Belfield. Activism by the AI Community: Analysing Recent Achievements and Future Prospects. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 15–21, New York NY USA, February 2020. ACM. ISBN 978-1-4503-7110-0. doi: 10.1145/3375627.3375814. 4.1, 4.6.4, 6.1.2

[30] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021. 1.2, 1.3, 3.4.3

[31] Garfield Benjamin. #FuckTheAlgorithm: Algorithmic imaginaries and political resistance. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 46–57, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533072. 4.1

[32] Cynthia L. Bennett and Daniela K. Rosner. The Promise of Empathy: Design, Disability, and Knowing the "Other". In *Conference on Human Factors in Computing Systems*, pages 1–13, Glasgow Scotland Uk, May 2019. ACM. ISBN 978-1-4503-5970-2. doi: 10.1145/3290605.3300528. 3.3.2

[33] Marco Berti and Ace V. Simpson. The Dark Side of Organizational Paradoxes: The Dynamics of Disempowerment. *Academy of Management Review*, 46(2):252–274, April 2021. ISSN 0363-7425, 1930-3807. doi: 10.5465/amr.2017.0208. URL `http://journals.aom.org/doi/10.5465/amr.2017.0208`. 1.4, 1.4.1, 4.1, 4.6.1

[34] Justin B Biddle. On predicting recidivism: epistemic risk, tradeoffs, and values in machine learning. *Canadian Journal of Philosophy*, pages 1–21, 2020. 1.3, 2.4.2

[35] Sam Biddle and Ryan Devereaux. Peter Thiel's Palantir Used to Bust Migrant Children's Relatives. https://theintercept.com/2019/05/02/peter-thiels-palantir-was-used-to-bust-hundreds-of-relatives-of-migrant-children-new-documents-show/, May 2019. 4.6.3

[36] 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. *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020. 5.6

[37] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. The Values Encoded in Machine Learning Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 173–184, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533083. 1.2

[38] Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. The Forgotten Margins of AI Ethics. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 948–958, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533157. 1.4, 4.1, 5.5, 6.1.2

[39] Sudershan Boovaraghavan, Chen Chen, Anurag Maravi, Mike Czapik, Yang Zhang, Chris Harrison, and Yuvraj Agarwal. Mites: Design and Deployment of a General-Purpose Sensing Infrastructure for Buildings. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(1):2:1–2:32, March 2023. doi: 10.1145/3580865. (document)

[40] Jason Borenstein and Ayanna Howard. Emerging challenges in AI and the need for AI ethics education. *AI and Ethics*, 1(1):61–65, February 2021. ISSN 2730-5953, 2730-5961. doi: 10.1007/s43681-020-00002-7. 4.6.4

[41] Nick Bostrom. Strategic implications of openness in ai development. *Global policy*, 8(2): 135–148, 2017. 2.1, 2.4.2

[42] Karen L. Boyd. Datasheets for datasets help ml engineers notice and understand ethical issues in training data. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), oct 2021. doi:

10.1145/3479582. URL https://doi.org/10.1145/3479582. 5.1

[43] Virginia Braun and Victoria Clarke. Thematic analysis. 2012. 4.2

[44] Enda Brophy. System Error: Labour Precarity and Collective Organizing at Microsoft. *Canadian Journal of Communication*, 31(3):619–638, October 2006. ISSN 0705-3657. doi: 10.22230/cjc.2006v31n3a1767. 4.6.2

[45] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023. 1.2

[46] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018. 1.2, 3.3.2

[47] Marcus Burkhardt. Mapping the democratization of ai on github. a first approach. 2019. 2.3.3

[48] Margaret Burnett, Simone Stumpf, Jamie Macbeth, Stephann Makri, Laura Beckwith, Irwin Kwan, Anicia Peters, and William Jernigan. GenderMag: A Method for Evaluating Software's Gender Inclusiveness. *Interacting with Computers*, 28(6):760–787, November 2016. ISSN 0953-5438, 1873-7951. doi: 10.1093/iwc/iwv046. 4.6.2

[49] Michel Callon. Introduction: the embeddedness of economic markets in economics. *The sociological review*, 46(1_suppl):1–57, 1998. 3.2

[50] Michael Carolan. Acting like an algorithm: digital farming platforms and the trajectories they (need not) lock-in. *Agriculture and Human Values*, 37(4):1041–1053, December 2020. ISSN 0889-048X, 1572-8366. doi: 10.1007/s10460-020-10032-w. URL https://link.springer.com/10.1007/s10460-020-10032-w. 3.1

[51] Stephanie Russo Carroll, Ibrahim Garba, Oscar L Figueroa-Rodríguez, Jarita Holbrook,

Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, et al. The care principles for indigenous data governance. 2020. 3.2, 3.4.3, 4.6.1

[52] Karin Knorr Cetina. *Epistemic cultures: How the sciences make knowledge*. Harvard University Press, 1999. 3.1

[53] Daniel Chandler. Technological or media determinism, 1995. 2.3.1

[54] Kelley Changfong-Hagen. " don't be evil": Collective action and employee prosocial activism. *HRLR Online*, 5:188, 2020. 4.6.4

[55] Kasia S Chmielinski, Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and Yue Chelsea Qiu. The dataset nutrition label (2nd gen): Leveraging context to mitigate harms in artificial intelligence. *arXiv preprint arXiv:2201.03954*, 2022. 3.4

[56] Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, 2020. 2.4.2

[57] Lars Thøger Christensen, Mette Morsing, and Ole Thyssen. License to Critique: A Communication Perspective on Sustainability Standards. *Business Ethics Quarterly*, 27(2): 239–262, April 2017. ISSN 1052-150X, 2153-3326. doi: 10.1017/beq.2016.66. 1.4.3, 5.1, 5.3.2, 5.5.1, 5.5.2, 5.6, 6.1.2, 6.2

[58] Peter Cihon, Jonas Schuett, and Seth D. Baum. Corporate Governance of Artificial Intelligence in the Public Interest. *Information*, 12(7):275, July 2021. ISSN 2078-2489. doi: 10.3390/info12070275. 4.1, 4.6, 4.6.4, 6.3, 6.4

[59] Victoria Clarke and Virginia Braun. Thematic analysis: a practical guide. *Thematic Analysis*, pages 1–100, 2021. 5.2.4

[60] CMU No Tech For ICE. Carnegie Mellon Dis-Orientation Guide. https://drive.google.com/file/u/2/d/1pS67e1Qf_jCzhdlhObk4v8JSuRp7OlJY/view?usp=sharing&usp=emb 2020. (document)

[61] Devin Coldewey. The riaa is coming for the youtube downloaders. *Tech Crunch*, October 2020. Accessed: 2021-07-1. 2.3.1

[62] E Gabriella Coleman. *Coding freedom: The ethics and aesthetics of hacking*. Princeton University Press, 2012. 2.1, 2.3.1, 2.3.3, 2.4.1, 3.3.1, 6.4

[63] Kate Conger and Daisuke Wakabayashi. Google employees say they faced retaliation after organizing walkout. *The New York Times*, April 2019. Accessed: 2023-06-27. 1.3, 5.1

[64] Danish Contractor, Daniel McDuff, Julia Katherine Haines, Jenny Lee, Christopher Hines, Brent Hecht, Nicholas Vincent, and Hanlin Li. Behavioral use licensing for responsible ai. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 778–788, 2022. 6.1.1

[65] Sasha Costanza-Chock. *Design justice: Community-led practices to build the worlds we need*. The MIT Press, 2020. 3.4.3

[66] Henriette Cramer, Jean Garcia-Gathright, Sravana Reddy, Aaron Springer, and Romain Takeo Bouyer. Translation, tracks & data: an algorithmic bias effort in practice. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2019. 2.4.3

[67] Henriette Cramer, Jenn Wortman Vaughan, Ken Holstein, Hanna Wallach, Jean Garcia-Gathright, Hal Daumé III, Miroslav Dudík, and Sravana Reddy. Challenges of incorporating algorithmic fairness into industry practice. *FAT* Tutorial*, 2019. 2.1

[68] Kate Crawford and Vladan Joler. Anatomy of an ai system. *Retrieved September*, 18: 2018, 2018. 1.1, 1.2, 6.1.1

[69] Jay Cunningham, Gabrielle Benabdallah, Daniela Rosner, and Alex Taylor. On the grounds of solutionism: Ontologies of blackness and hci. *ACM Transactions on Computer-Human Interaction*, 30(2):1–17, 2023. 5.3.2

[70] Laura Dabbish, Colleen Stuart, Jason Tsay, and Jim Herbsleb. Social coding in github:

transparency and collaboration in an open software repository. In *ACM 2012 conference on computer supported cooperative work*, pages 1277–1286, 2012. 2.2.1, 2.3.2

[71] Sherae L Daniel, Ting-Ting Rachel Chung, and Pratyush Nidhi Sharma. The impact of anonymous peripheral contributions on open source software development. *AIS Transactions on Human-Computer Interaction*, 12(3):146–171, 2020. 2.1

[72] Tom Davidson. Pittsburgh City Council told facial recognition bill should be stronger. *TribLIVE*, September 2020. (document)

[73] Jenny L Davis. *How artifacts afford: The power and politics of everyday things*. MIT Press, 2020. 4.6.4

[74] Tim De Chant. Google hired union-busting consultants to convince employees "unions suck" — Ars Technica. https://arstechnica.com/tech-policy/2022/01/google-hired-union-busting-consultants-to-convince-employees-unions-suck/, 1/11/2022, 12:46 PM. 6.4

[75] Stanley Deetz. *Democracy in an age of corporate colonization: Developments in communication and the politics of everyday life*. SUNY press, 1992. 5.1

[76] Wesley Hanwen Deng, Nur Yildirim, Monica Chang, Motahhare Eslami, Kenneth Holstein, and Michael Madaio. Investigating Practices and Opportunities for Cross-functional Collaboration around AI Fairness in Industry Practice. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 705–716, Chicago IL USA, June 2023. ACM. ISBN 9798400701924. doi: 10.1145/3593013.3594037. 6.1.1, 6.1.2, 6.2

[77] Edsger W Dijkstra. On the role of scientific thought. In *Selected writings on computing: a personal perspective*, pages 60–66. Springer, 1982. 3.1

[78] Tulsee Doshi and Andrew Zaldivar. Responsible ai with tensorflow. *TensorFlow Blog*, June 2020. Accessed: 2021-07-7. 2.4.3, 6.2

[79] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580, January 2018. ISSN 2375-2548. doi: 10.1126/sciadv.

aao5580. 4.1

[80] Mengnan Du, Fan Yang, Na Zou, and Xia Hu. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, 36(4):25–34, 2020. 1.3

[81] Joseph Dumit. Writing the Implosion: Teaching the World One Thing at a Time. *Cultural Anthropology*, 29(2):344–362, May 2014. ISSN 1548-1360, 0886-7356. doi: 10.14506/ca29.2.09. 6.1.1

[82] Anthony Dunne and Fiona Raby. *Speculative Everything: Design, Fiction, and Social Dreaming*. MIT Press, December 2013. ISBN 978-0-262-01984-2. Google-Books-ID: 9gQyAgAAQBAJ. 5.1

[83] Jean Macchiaroli Eggen and John G Culhane. Gun torts: Defining a cause of action for victims in suits against gun manufacturers. *NCL Rev.*, 81:115, 2002. 2.4.3

[84] Joel Feinberg. Collective responsibility. *The Journal of Philosophy*, 65(21):674–688, 1968. 2.4.2

[85] Andrew Feller, Dan Shunk, and Tom Callarman. Value chains versus supply chains. *BP trends*, 1:1–7, 2006. 3.4.2

[86] Avi Feller, Emma Pierson, Sam Corbett-Davies, and Sharad Goel. A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear. *The Washington Post*, 17, 2016. 4.1

[87] Heike Felzmann, Eduard Fosch-Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. Towards transparency by design for artificial intelligence. *Science and Engineering Ethics*, 26(6):3333–3361, 2020. 1.3

[88] James Ferguson. *The anti-politics machine:" development," depoliticization, and bureaucratic power in Lesotho*. U of Minnesota Press, 1994. 3.4.2

[89] Casey Fiesler and Natalie Garrett. Ethical tech starts with addressing ethical debt. 2020. 3.1, 3.2, 6.3

[90] Casey Fiesler, Natalie Garrett, and Nathan Beard. What Do We Teach When We Teach Tech Ethics?: A Syllabi Analysis. In *51st ACM Technical Symposium on Computer Science Education*, pages 289–295, Portland OR USA, February 2020. ACM. ISBN 978-1-4503-6793-6. doi: 10.1145/3328778.3366825. 4.6.2, 4.6.3, 4.6.4, 6.3

[91] Peter Fleming and André Spicer. Power in Management and Organization Science. *Academy of Management Annals*, 8(1):237–298, January 2014. ISSN 1941-6520, 1941-6067. doi: 10.5465/19416520.2014.875671. URL `http://journals.aom.org/doi/10.5465/19416520.2014.875671`. 1.4, 1.4.1, 4.1, 4.6.1, 6.1.2

[92] Michel Foucault. *Power/knowledge: Selected interviews and other writings, 1972-1977*. Vintage, 1980. 1.4

[93] Sarah E. Fox, Vera Khovanskaya, Clara Crivellaro, Niloufar Salehi, Lynn Dombrowski, Chinmay Kulkarni, Lilly Irani, and Jodi Forlizzi. Worker-Centered Design: Expanding HCI Methods for Supporting Labor. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–8, Honolulu HI USA, April 2020. ACM. ISBN 978-1-4503-6819-3. doi: 10.1145/3334480.3375157. 6.1.2

[94] Batya Friedman. Value-sensitive design. *interactions*, 3(6):16–23, 1996. 3.4.2

[95] Hana Frluckaj, Laura Dabbish, David Gray Widder, Huilian Sophie Qiu, and James D. Herbsleb. Gender and Participation in Open Source Software Development. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–31, November 2022. ISSN 2573-0142. doi: 10.1145/3555190. 6.1.2

[96] Ben Gansky and Sean McDonald. Counterfacctual: How facct undermines its organizing principles. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1982–1992, 2022. 1.3, 1.4, 3.1, 4.6.3, 4.6.4, 5.5, 6.1.1, 6.1.2, 6.2

[97] Marta Garnelo and Murray Shanahan. Reconciling deep learning with symbolic artificial intelligence: Representing objects and relations. *Current Opinion in Behavioral Sciences*,

29:17–23, October 2019. ISSN 2352-1546. doi: 10.1016/j.cobeha.2018.12.010. 1.2

[98] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. 1.3, 3.4, 4.1, 4.6.2

[99] R Stuart Geiger. Summary analysis of the 2017 github open source survey. *arXiv preprint arXiv:1706.02777*, 2017. 2.2.2

[100] Erving Goffman. *The presentation of self in everyday life*. Anchor, 1959. 3.3.2

[101] Marcela F González. Precarity for the global talent: The impact of visa policies on high-skilled immigrants' work in the united states. *International Migration*, 60(2):193–207, 2022. 4.5.1

[102] Google. Consultation on the eu ai act proposal: Google's submission, 15 July 2021. URL `https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2662492_en`. 6.4

[103] Chandell Gosse and Jacquelyn Burkell. Politics and porn: how news media characterizes problems presented by deepfakes. *Critical Studies in Media Communication*, 37(5):497–511, 2020. 2.2.1

[104] DW Gotterbarn, Bo Brinkman, Catherine Flick, Michael S Kirkpatrick, Keith Miller, Kate Vazansky, and Marty J Wolf. Acm code of ethics and professional conduct. 2018. 4.1

[105] Colin Grant. Friedman fallacies. *Journal of Business Ethics*, 10(12):907–914, 1991. 3.2

[106] Colin M Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L Toombs. The dark (patterns) side of ux design. In *2018 CHI conference on human factors in computing systems*, pages 1–14, 2018. 4.3.1

[107] Mary Gray. The banality of scale: A theory on the limits of modeling bias and

fairness frameworks for social justice. Conference on Neural Information Processing Systems (NeurIPS), 2021. URL https://neurips.cc/virtual/2021/invited-talk/22281. 3.3.1, 3.4.3

[108] Daniel Greene, Anna Lauren Hoffmann, and Luke Stark. Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In *52nd Hawaii international conference on system sciences*, 2019. 1.1, 1.3, 1.4.3, 2.1, 2.2.1, 2.4.2, 3.1, 4.1, 4.6.1, 4.6.2, 4.6.4, 4.7, 5.1, 5.5.2, 5.6, 6.1.1, 6.1.2, 6.2, 6.4

[109] Frances S Grodzinsky, Keith Miller, and Marty J Wolf. Ethical issues in open source software. *Journal of Information, Communication and Ethics in Society*, 2003. 2.4.2

[110] Eileen Guo and Tate Ryan-Mosley. Computer scientists designing the future can't agree on what privacy means. https://www.technologyreview.com/2023/04/03/1070665/cmu-university-privacy-battle-smart-building-sensors-mites/, April 2023. (document)

[111] Philip J Guo, Thomas Zimmermann, Nachiappan Nagappan, and Brendan Murphy. Characterizing and predicting which bugs get fixed: an empirical study of microsoft windows. In *32Nd ACM/IEEE International Conference on Software Engineering-Volume 1*, pages 495–504, 2010. 4.6.2

[112] Seda Gürses and Joris van Hoboken. *Privacy after the Agile Turn*, page 579–601. Cambridge Law Handbooks. Cambridge University Press, 2018. doi: 10.1017/9781316831960.032. 3.2

[113] Alex Hanna and Tina M. Park. Against Scale: Provocations and Resistances to Scale Thinking, November 2020. 3.1, 3.3.1, 3.4.3

[114] Karen Hao. In 2020, let's stop ai ethics-washing and actually do something. *MIT Technology Review*, 27(December):2019, 2019. 4.1

[115] Donna J Haraway. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Simians, cyborgs, and women: The reinvention of nature*, pages

183–201, 1991. 1.4.2, 1, 3.2, 5.5.1

[116] Thomas A. Hemphill. 'Techlash', responsible innovation, and the self-regulatory organization. *Journal of Responsible Innovation*, 6(2):240–247, May 2019. ISSN 2329-9460, 2329-9037. doi: 10.1080/23299460.2019.1602817. 4.1

[117] Guido Hertel, Sven Niedner, and Stefanie Herrmann. Motivation of software developers in open source projects: an internet-based survey of contributors to the linux kernel. *Research policy*, 32(7):1159–1177, 2003. 2.3.2

[118] Ralph Hertwig, Carola Fanselow, and Ulrich Hoffrage. Hindsight bias: How knowledge and heuristics affect our reconstruction of the past. *Memory*, 2003. 4.2

[119] Bill Hibbard. Open source ai. *Frontiers in Artificial Intelligence and Applications*, 171: 473, 2008. 2.4.2

[120] Claudia Hilderbrand, Christopher Perdriau, Lara Letaw, Jillian Emard, Zoe Steine-Hanson, Margaret Burnett, and Anita Sarma. Engineering gender-inclusivity into software: Ten teams' tales from the trenches. In *ACM/IEEE 42nd International Conference on Software Engineering*, pages 433–444, Seoul South Korea, June 2020. ACM. ISBN 978-1-4503-7121-6. doi: 10.1145/3377811.3380371. 4.6.2

[121] Matthew Hockenberry. Redirected entanglements in the digital supply chain. *Cultural studies*, 35(4-5):641–662, 2021. 3.1, 3.2

[122] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *2019 CHI conference on human factors in computing systems*, pages 1–16, 2019. 2.1, 3.4, 5.1, 6.1.1, 6.1.2, 6.4

[123] Daning Hu, J Leon Zhao, and Jiesi Cheng. Reputation management in an open source developer social network: An empirical study on determinants of positive evaluations. *Decision Support Systems*, 53(3):526–533, 2012. 2.3.2

[124] George P. Huber and Kyle Lewis. Cross-Understanding: Implications for Group Cognition and Performance. *The Academy of Management Review*, 35(1):6–26, 2010. ISSN 0363-7425. URL https://www.jstor.org/stable/27760038. Publisher: Academy of Management. 5.5.1

[125] Lee Humphreys. Reframing Social Groups, Closure, and Stabilization in the Social Construction of Technology. *Social Epistemology*, 19(2-3):231–253, January 2005. ISSN 0269-1728. doi: 10.1080/02691720500145449. URL https://doi.org/10.1080/02691720500145449. Publisher: Routledge _eprint: https://doi.org/10.1080/02691720500145449. 1.4.3, 5.5.2, 6.1.1

[126] Don Ihde. Technology and the lifeworld: From garden to earth. 1990. 2.3.1, 2.4.2

[127] Lilly Irani. *Chasing Innovation: Making Entrepreneurial Citizens in Modern India*. Princeton University Press, March 2019. ISBN 978-0-691-18944-4. doi: 10.1515/9780691189444. 4.6.2, 6.1.2

[128] 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*, pages 611–620, 2013. 1.2

[129] William Isaac and Karen Hao. Keynote interview: Karen hao in conversation with william isaac. https://www.youtube.com/watch?v=9u-62Ijtb1I, Jun 2022. 3.4.2

[130] Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. How Different Groups Prioritize Ethical Values for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 310–323, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533097. 4.1

[131] James Vincent. Google employee who helped lead protests leaves company - The Verge. July 2019. URL https://www.theverge.com/2019/7/16/20695964/

google-protest-leader-meredith-whittaker-leaves-company. 1.3

[132] Niranjan S. Janardhanan, Kyle Lewis, Rhonda K. Reger, and Cynthia K. Stevens. Getting to Know You: Motivating Cross-Understanding for Improved Team and Individual Performance. *Organization Science*, 31(1):103–118, January 2020. ISSN 1047-7039. doi: 10.1287/orsc.2019.1324. URL https://pubsonline.informs.org/doi/abs/10.1287/orsc.2019.1324. Publisher: INFORMS. 5.5.1

[133] Sue Curry Jansen and Brian Martin. The streisand effect and censorship backfire. 2015. 2.3.1

[134] Seyyed Ahmad Javadi, Chris Norval, Richard Cloete, and Jatinder Singh. Monitoring ai services for misuse. In *2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 597–607, 2021. 3.4.2, 6.2

[135] Colin Jerolmack and Shamus Khan. Talk is cheap: Ethnography and the attitudinal fallacy. *Sociological methods & research*, 43(2):178–209, 2014. 2.2.2

[136] Jiji. The riaa is coming for the youtube downloaders. *The Japan Times*, October 2020. Accessed: 2021-07-1. 2.3.1

[137] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019. 1.1, 1.3, 1.4.3, 2.1, 2.3.3, 3.1, 4.1, 4.6.1, 4.6.2, 4.7, 5.1, 6.4

[138] Khari Johnson. Ai ethics is all about power. *Venture Beat*, 1, 2019. 5.5

[139] Jason Karaian and Lora Kelley. How big tech layoffs stack up with the rest of their work forces. https://www.nytimes.com/2023/01/21/business/tech-layoffs.html, January 2023. 4.6

[140] Christopher M Kelty. *Two bits: The cultural significance of free software*. Duke University Press, 2008. 2.3.1, 2.3.1, 2.4.1

[141] Chris F Kemerer. Software complexity and software maintenance: A survey of empirical

research. *Annals of Software Engineering*, 1(1):1–22, 1995. 3.1

[142] 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*, pages 1–11, 2019. 1.1, 1.3, 2.1, 2.4.2, 4.1, 4.6.1, 4.7, 5.1, 5.5.2, 6.1.2, 6.2, 6.4

[143] Jan Kietzmann, Linda W Lee, Ian P McCarthy, and Tim C Kietzmann. Deepfakes: Trick or treat? *Business Horizons*, 63(2):135–146, 2020. 2.1, 2.2.1

[144] Miryung Kim, Thomas Zimmermann, and Nachiappan Nagappan. An empirical study of refactoringchallenges and benefits at microsoft. *IEEE Transactions on Software Engineering*, 40(7):633–649, 2014. 3.1

[145] P. M. Krafft, Meg Young, Michael Katell, Karen Huang, and Ghislain Bugingo. Defining AI in Policy versus Practice. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 72–78, New York NY USA, February 2020. ACM. ISBN 978-1-4503-7110-0. doi: 10.1145/3375627.3375835. URL `https://dl.acm.org/doi/10.1145/3375627.3375835`. 4.6.4

[146] Logan Kugler. The unionization of technology companies. *Communications of the ACM*, 64(8):18–20, 2021. 4.4.4, 4.6.4

[147] Nick Lally. "it makes almost no difference which algorithm you use": on the modularity of predictive policing. *Urban Geography*, pages 1–19, 2021. 3.1

[148] Bruno Latour. *We have never been modern*. Harvard university press, 1993. 3.1

[149] Bruno Latour. On technical mediation. 1994. 2.3.1, 2.4.2

[150] Bruno Latour, Wiebe E Bijker, and John Law. Shaping technology/building society: Studies in sociotechnical change. *W. Bijker & J. Law (Eds.)*, pages 225–258, 1992. 2.3.1

[151] Bruno Latour et al. *Pandora's hope: essays on the reality of science studies*. Harvard university press, 1999. 3.1

[152] Thomas B. Lawrence. Power, Institutions and Organizations. In *The SAGE Handbook of Organizational Institutionalism*, pages 170–197. SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom, 2008. ISBN 978-1-4129-3123-6 978-1-84920-038-7. doi: 10.4135/9781849200387.n7. URL `https://sk.sagepub.com/reference/hdbk_orginstitution/n7.xml`. 4.1

[153] Thomas B Lawrence and Sean Buchanan. Power, institutions and organizations. *The Sage handbook of organizational institutionalism*, pages 477–506, 2017. 1.4, 1.4.1, 4.6.1

[154] Jennifer Lee, Meg Young, PM Krafft, and Michael A Katell. Power and technology: Who gets to make the decisions? *Interactions*, 28(1):38–46, 2020. 5.5

[155] Jennifer Lee, Meg Young, P. M. Krafft, and Michael A. Katell. Power and technology: Who gets to make the decisions? *Interactions*, 28(1):38–46, January 2021. ISSN 1072-5520, 1558-3449. doi: 10.1145/3442420. 1.4, 6.1.2

[156] Michelle Seng Ah Lee and Jat Singh. The landscape and gaps in open source fairness toolkits. In *2021 CHI conference on human factors in computing systems*, pages 1–13, 2021. 4.6.2

[157] Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. Working with machines: The impact of algorithmic and data-driven management on human workers. In *33rd annual ACM conference on human factors in computing systems*, pages 1603–1612, 2015. 4.1, 4.3.1

[158] Max Liboiron. Pollution is colonialism. In *Pollution Is Colonialism*. Duke University Press, 2021. 3.3.1

[159] Chien-Te Lin. All about the human: A Buddhist take on AI ethics. *Business Ethics, the Environment & Responsibility*, 32(3):1113–1122, 2023. ISSN 2694-6424. doi: 10.1111/beer.12547. 1.3

[160] Yvonna S Lincoln, Susan A Lynham, and Egon G Guba. Paradigmatic controversies,

contradictions, and emerging confluences, revisited. *The Sage handbook of qualitative research*, 4:97–128, 2011. 2.2.2, 5.2.4

[161] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018. 2.1, 4.1

[162] Alan MacCormack, John Rusnak, and Carliss Y. Baldwin. Exploring the Structure of Complex Software Designs: An Empirical Study of Open Source and Proprietary Code. *Management Science*, 52(7):1015–1030, July 2006. ISSN 0025-1909, 1526-5501. doi: 10.1287/mnsc.1060.0552. URL `http://pubsonline.informs.org/doi/10.1287/mnsc.1060.0552`. 3.1

[163] Kathleen M MacQueen, Eleanor McLellan, Kelly Kay, and Bobby Milstein. Codebook development for team-based qualitative analysis. *Cam Journal*, 10(2):31–36, 1998. 2.2.2

[164] 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. *ACM Conference on Human-Computer Interaction*, 6 (CSCW1):1–26, 2022. 3.3.1, 4.1, 5.1

[165] 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 *2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020. 1.1, 1.3, 2.1, 3.1, 3.4, 4.1, 4.6.2, 4.7, 5.1, 5.4.2, 5.5.1, 5.6, 6.1.1, 6.1.2, 6.2, 6.4

[166] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Microsoft ai fairness checklist. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4t6dA, 2020. 5.5.2

[167] Sophie Maddocks. 'a deepfake porn plot intended to silence me': exploring continuities between pornographic and 'political'deep fakes. *Porn Studies*, 7(4):415–423, 2020. 2.2.1

[168] James W Malazita and Korryn Resetar. Infrastructures of abstraction: how computer science education produces anti-political subjects. *Digital Creativity*, 30(4):300–312, 2019. 3.1

[169] Travis Mandel, Jahnu Best, Randall H. Tanaka, Hiram Temple, Chansen Haili, Sebastian J. Carter, Kayla Schlechtinger, and Roy Szeto. Using the crowd to prevent harmful ai behavior. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2), oct 2020. doi: 10.1145/3415168. URL `https://doi.org/10.1145/3415168`. 5.1

[170] Noëmi Manders-Huits and Michael Zimmer. Values and pragmatic action: The challenges of introducing ethical intelligence in technical design communities. *The International Review of Information Ethics*, 10:37–44, 2009. 2.1

[171] Jennifer Mankoff, Jennifer A Rode, and Haakon Faste. Looking past yesterday's tomorrow: using futures studies methods to extend the research horizon. page 10, April 2013. 5.1

[172] James Manyika. Getting AI Right: Introductory Notes on AI & Society. *Daedalus*, 151 (2):5–27, May 2022. ISSN 0011-5266, 1548-6192. doi: 10.1162/daed_e_01897. 1.2

[173] Gary Marcus. The Sparks of AGI? Or the End of Science? https://cacm.acm.org/blogs/blog-cacm/271354-the-sparks-of-agi-or-the-end-of-science/fulltext, March 2023. 1.2

[174] Jennifer Marlow and Laura Dabbish. Activity traces and signals in software developer recruitment and hiring. In *2013 conference on Computer supported cooperative work*, pages 145–156, 2013. 2.3.2

[175] Nikolas Martelaro and Wendy Ju. What could go wrong? exploring the downsides of autonomous vehicles. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 99–101, 2020. 5.5.1

[176] Martin Gilesarchive. Artificial intelligence is often overhyped—and

here's why that's dangerous | MIT Technology Review, September 2018. URL https://web.archive.org/web/20230311145742/https://www.technologyreview.com/2018/09/13/240156/artificial-intelligence-is-often-overhypedand-heres-why-thats-dangerou 1.2

[177] Andrew McNamara, Justin Smith, and Emerson Murphy-Hill. Does acm's code of ethics change ethical decision making in software development? In *2018 26th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering*, pages 729–733, 2018. 4.1

[178] Tara McPherson. *Feminist in a Software Lab: Difference+ Design*, volume 6. Harvard University Press, 2018. 3.1

[179] Jacob Metcalf, Emanuel Moss, et al. Owning ethics: Corporate logics, silicon valley, and the institutionalization of ethics. *Social Research: An International Quarterly*, 86(2): 449–476, 2019. 1.3, 2.1, 3.1, 3.3.2, 4.1, 4.7, 5.1, 5.5.2, 6.1.2

[180] Cade Metz. 'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead. *The New York Times*, May 2023. ISSN 0362-4331. URL https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html. 1.3, 5.1

[181] Cade Metz and Daisuke Wakabayashi. Google Researcher Says She Was Fired Over Paper Highlighting Bias in A.I. *The New York Times*, December 2020. ISSN 0362-4331. 1.3, 5.1

[182] Debra E Meyerson. *Rocking the boat: How tempered radicals effect change without making trouble*. Harvard Business Review Press, 2008. 4.4.2, 4.6.1

[183] Matthew B Miles and A Michael Huberman. *Qualitative data analysis: An expanded sourcebook*. sage, 1994. 2.2.2

[184] 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 *ACM conference on fairness, accountability, and transparency*, pages 220–229, 2019. 1.3, 3.4, 4.1, 4.6.2

[185] Brent Mittelstadt. Principles alone cannot guarantee ethical ai. *Nature Machine Intelligence*, 1(11):501–507, 2019. 1.1, 1.3, 2.1, 4.1

[186] Deirdre K Mulligan and Kenneth A Bamberger. Procurement as policy: Administrative process for machine learning. *Berkeley Tech. LJ*, 34:773, 2019. 6.2

[187] Lewis Mumford. *The pentagon of power*, volume 2. Houghton Mifflin Harcourt P, 1970. 2.3.1

[188] Robin Murphy and David D Woods. Beyond asimov: the three laws of responsible robotics. *IEEE intelligent systems*, 24(4):14–20, 2009. 2.2.1

[189] Anton J Nederhof. Methods of coping with social desirability bias: A review. *European journal of social psychology*, 1985. 4.2

[190] Nataliya Nedzhvetskaya and J. S. Tan. The Role of Workers in AI Ethics and Governance. In Justin B. Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew M. Young, and Baobao Zhang, editors, *The Oxford Handbook of AI Governance*, pages C68.S1–C68.N14. Oxford University Press, first edition, August 2022. ISBN 978-0-19-757932-9 978-0-19-757935-0. doi: 10.1093/oxfordhb/9780197579329.013.68. 4.1, 4.6.4

[191] Nataliya Nedzhvetskaya, JS Tan, Hyatt Dirbas, and Wynnie Chan. Collective Action in Tech. https://data.collectiveaction.tech/, September 2022. 4.1, 4.6.1, 4.6.4

[192] Gina Neff. From Bad Users and Failed Uses to Responsible Technologies: A Call to Expand the AI Ethics Toolkit. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 5–6, New York NY USA, February 2020. ACM. ISBN 978-1-4503-7110-0. doi:

10.1145/3375627.3377141. 4.6.3

[193] Nils J. Nilsson. *The Quest for Artificial Intelligence*. Cambridge University Press, 1 edition, October 2009. ISBN 978-0-521-11639-8 978-0-521-12293-1 978-0-511-81934-6. doi: 10.1017/CBO9780511819346. URL https://www.cambridge.org/core/product/identifier/9780511819346/type/book. 1.2

[194] Helen Nissenbaum. Accountability in a computerized society. *Science and Engineering Ethics*, 2(1):25–42, March 1996. ISSN 1471-5546. doi: 10.1007/BF02639315. 3.1, 6.2

[195] Helen Nissenbaum. Computing and accountability. In *Computer Ethics*, pages 273–280. Routledge, 2017. 2.4.2

[196] Rodrigo Ochigame. How Big Tech Manipulates Academia to Avoid Regulation. https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/, December 2019. 1.4, 4.1, 4.6.2, 6.1.2

[197] Academy of Motion Picture Arts and Sciences. The 82nd academy awards (2010) nominees and winners. *The Academy of Motion Picture Arts and Sciences*, 2010. Accessed: 2022-004-25. 2.3.3

[198] Wonseok Oh and Sangyong Jeon. Membership herding and network stability in the open source community: The ising perspective. *Management science*, 53(7):1086–1101, 2007. 2.1

[199] Marc Ohm, Henrik Plate, Arnold Sykosch, and Michael Meier. Backstabber's knife collection: A review of open source software supply chain attacks. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 23–43. Springer, 2020. 2.4.3

[200] OpenAI. GPT-4 Technical Report, March 2023. URL http://arxiv.org/abs/2303.08774. arXiv:2303.08774 [cs]. 1.2

[201] Will Orr and Jenny L Davis. Attributions of ethical responsibility by artificial intelligence

practitioners. *Information, Communication & Society*, 23(5):719–735, 2020. 1.3, 2.1, 2.2.1, 2.3.1, 2.4.2, 2.4.3, 3.1, 4.1, 6.1.2

[202] Arnold Pacey. *The culture of technology*. MIT press, 1983. 2.3.1

[203] Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. Designing fair ai in human resource management: Understanding tensions surrounding algorithmic evaluation and envisioning stakeholder-centered solutions. In *CHI Conference on Human Factors in Computing Systems*, pages 1–22, 2022. 4.1

[204] David L Parnas. On the criteria to be used in decomposing systems into modules. In *Pioneers and Their Contributions to Software Engineering*, pages 479–498. Springer, 1972. 3.1

[205] Jussi Pasanen. Human-centred design considered harmful. 2019. Accessed: 2022-05-2. 3.3.2

[206] Bruce Perens. The emerging economic paradigm of open source. *First Monday*, 2005. 2.3.1

[207] Peter Singer. Ethics, May 2023. URL `https://www.britannica.com/topic/ethics-philosophy`. 1.3

[208] Sundar Pichai. Ai at google: our principles. *Google Blog*, June 2018. Accessed: 2023-06-19. 6.2

[209] Stuart Plattner. *Economic anthropology*. Stanford University Press, 1989. 3.2

[210] Louis P Pojman and James Fieser. Introduction to philosophy: classical and contemporary readings. 2004. 1.3

[211] Miriam Posner. See no evil. *Logic Magazine*, 2018. 3.2

[212] Oxford University Press. agency, n., . URL `https://www.oed.com/view/Entry/3851`. 1.4

[213] Oxford University Press. power, n., . URL https://www.oed.com/view/Entry/149167. 1.4

[214] Robert N Proctor and Londa Schiebinger. Agnotology: The making and unmaking of ignorance. 2008. 3.1

[215] Hatim A. Rahman. The Invisible Cage: Workers' Reactivity to Opaque Algorithmic Evaluations. *Administrative Science Quarterly*, 66(4):945–988, December 2021. ISSN 0001-8392. doi: 10.1177/00018392211010118. 4.1

[216] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. Where responsible ai meets reality: Practitioner perspectives on enablers for shifting organizational practices. *In 24th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 2020. 4.1, 4.6.2, 4.6.4, 5.1

[217] 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 *2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Yokohama Japan, May 2021. ACM. ISBN 978-1-4503-8096-6. doi: 10.1145/3411764.3445604. 4.6.2

[218] Marc J Riemer. Communication skills for the 21st century engineer. *Global J. of Engng. Educ*, 11(1):89–100, 2007. 4.2

[219] Carol Righi, Janice James, Michael Beasley, Donald L Day, Jean E Fox, Jennifer Gieber, Chris Howe, and Laconya Ruby. Card sort analysis best practices. *Journal of Usability Studies*, 8(3):69–89, 2013. 2.2.2

[220] Sarah Roberts. Supply chain specific? understanding the patchy success of ethical sourcing initiatives. *Journal of business ethics*, 44(2):159–170, 2003. 2.4.3

[221] Cynthia Rudin, Caroline Wang, and Beau Coker. The age of secrecy and unfairness in recidivism prediction. *arXiv preprint arXiv:1811.00731*, 2018. 2.4.2

[222] John Ruggie. 2011. 2.4.3

[223] Henrik Skaug Sætra, Mark Coeckelbergh, and John Danaher. The AI ethicist's dilemma: Fighting Big Tech by supporting Big Tech. *AI and Ethics*, 2(1):15–27, February 2022. ISSN 2730-5953, 2730-5961. doi: 10.1007/s43681-021-00123-7. 4.6.4

[224] Niloufar Salehi, Lilly C Irani, Michael S Bernstein, Ali Alkhatib, Eva Ogbe, and Kristy Milland. 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*, pages 1621–1630, 2015. 4.1, 6.1.2

[225] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021. 1.2, 3.3.1, 3.4

[226] Barry R Schneider. Nuclear proliferation and counter-proliferation: Policy issues and debates. *Mershon International Studies Review*, 38(Supplement_2):209–234, 1994. 2.3.1

[227] James C Scott. *Weapons of the weak: Everyday forms of peasant resistance*. Yale University Press, 1985. 3.3.2

[228] James C Scott. *Domination and the arts of resistance: Hidden transcripts*. Yale university press, 1990. 1.4, 1.4.3, 3.3.2, 5.1, 5.4.1, 5.5, 5.5.1, 5.6, 6.1.2

[229] SCS Dean's PhD Advisory Committee. Towards Anti-Racist Change in the School of Computer Science. September 2020. (document)

[230] Evan Selinger. The philosophy of the technology of the gun. *The Atlantic*, 23, 2012. 2.3.1

[231] Mary Shaw. Modularity for the modern world: summary of invited keynote. In *tenth international conference on Aspect-oriented software development*, pages 1–6, 2011. 3.1

[232] Keith Simmons. Kant on Moral Worth. *History of Philosophy Quarterly*, 6(1):85–100, 1989. ISSN 0740-0675. 1.3

[233] Sean Sirur, Jason RC Nurse, and Helena Webb. Are we there yet? understanding the challenges faced in complying with the general data protection regulation (gdpr). In *2nd International Workshop on Multimedia Privacy and Security*, pages 88–95, 2018. 3.1

[234] Michael Skirpan, Maggie Oates, Daragh Byrne, Robert Cunningham, and Lorrie Faith Cranor. Is a privacy crisis experienced, a privacy crisis avoided? *Communications of the ACM*, 65(3):26–29, March 2022. ISSN 0001-0782, 1557-7317. doi: 10.1145/3512325. URL `https://dl.acm.org/doi/10.1145/3512325`. 5.1

[235] Reginald G Smart. Subject selection bias in psychological research. *Canadian Psychologist/Psychologie canadienne*, 1966. 4.2

[236] Irene Solaiman. The Gradient of Generative AI Release: Methods and Considerations. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 111–122, Chicago IL USA, June 2023. ACM. ISBN 9798400701924. doi: 10.1145/3593013.3593981. 6.1.1

[237] Richard Stallman. Why programs must not limit the freedom to run them. *The Free Software Foundation*, November 2016. Accessed: 2021-07-1. 2.1, 2.3.3, 6.1.1

[238] Richard Stallman et al. What is free software?, the free software definition, version 1.169. *The Free Software Foundation*, February 2021. Accessed: 2021-07-1. 2.1, 2.4.2, 6.2

[239] Luke Stark, Daniel Greene, and Anna Lauren Hoffmann. Critical perspectives on governance mechanisms for ai/ml systems. In *The Cultural Life of Machine Learning*, pages 257–280. Springer, 2021. 2.4.1

[240] Stefan Stieger and Anja S Göritz. Using instant messaging for internet-based interviews. *CyberPsychology & Behavior*, 9(5):552–559, 2006. 2.2.2

[241] Marilyn Strathern. An anthropological comment. *Virtual society?: Technology, cyberbole, reality*, page 302, 2002. 3.2

[242] Anselm Strauss and Juliet Corbin. *Basics of qualitative research*. Sage publications, 1990.

5.2.4

[243] Anselm Strauss and Juliet M Corbin. *Grounded theory in practice*. Sage, 1997. 2.2.2

[244] Eliza Strickland. Timnit gebru is building a slow ai movement. *IEEE Spectrum*, March 2022. Accessed: 2022-07-7. 3.4.3

[245] Norman Makoto Su, Amanda Lazar, and Lilly Irani. Critical affects: Tech work emotions amidst the techlash. *ACM Conference on Human-Computer Interaction*, 5(CSCW1):1–27, 2021. 4.6.4, 6.1.2

[246] Hariharan Subramonyam, Jane Im, Colleen Seifert, and Eytan Adar. Solving separation-of-concerns problems in collaborative design of human-ai systems through leaky abstractions. In *CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2022. 3.3.1

[247] Lucy Suchman. Located accountabilities in technology production. *Scandinavian journal of information systems*, 14(2):7, 2002. 1.3, 1.4.2, 3.1, 3.3.1, 5.5.1, 6.1.2

[248] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning Series. The MIT Press, Cambridge, Massachusetts, second edition edition, 2018. ISBN 978-0-262-03924-6. 1.2

[249] David R Thomas. A general inductive approach for qualitative data analysis. 2003. 2.2.2

[250] Suzanne L Thomas. Migration versus management: the global distribution of computer vision engineering work. In *2019 ACM/IEEE 14th International Conference on Global Software Engineering (ICGSE)*, pages 12–17. IEEE, 2019. 3.2, 3.2, 4.6.4

[251] M Tiemann and Open Source Initiative. History of the osi (open source initiative). *Retrieved February*, 4:2009, 2009. 2.4.1

[252] Marcus Tomalin, Bill Byrne, Shauna Concannon, Danielle Saunders, and Stefanie Ullmann. The practical ethics of bias reduction in machine translation: Why domain adaptation is better than data debiasing. *Ethics and Information Technology*, pages 1–15, 2021.

2.4.2

[253] Elizabeth C. Tomlinson. Stasis in the Shark Tank: Persuading an Audience of Funders to Act on Behalf of Entrepreneurs. *Journal of Business and Technical Communication*, March 2020. doi: 10.1177/1050651920910219. URL https://journals.sagepub.com/doi/10.1177/1050651920910219. Publisher: SAGE PublicationsSage CA: Los Angeles, CA. 5.5.2

[254] Ville Vakkuri, Kai-Kristian Kemell, Marianna Jantunen, and Pekka Abrahamsson. "This is Just a Prototype": How Ethics Are Ignored in Software Startup-Like Environments. In Viktoria Stray, Rashina Hoda, Maria Paasivaara, and Philippe Kruchten, editors, *Agile Processes in Software Engineering and Extreme Programming*, volume 383, pages 195–210. Springer International Publishing, Cham, 2020. ISBN 978-3-030-49391-2 978-3-030-49392-9. doi: 10.1007/978-3-030-49392-9_13. URL http://link.springer.com/10.1007/978-3-030-49392-9_13. Series Title: Lecture Notes in Business Information Processing. 4.1

[255] Shannon Vallor, Brian Green, and Irina Raicu. Ethics in technology practice. *The Markkula Center for Applied Ethics at Santa Clara University. https://www. scu. edu/ethics*, 2018. 3.4

[256] Joachim Van den Bergh and Dirk Deschoolmeester. Ethical decision making in ict: discussing the impact of an ethical code of conduct. *Communications of the IBIMA*, pages 1–10, 2010. 2.1

[257] Rama Adithya Varanasi and Nitesh Goyal. "it is currently hodgepodge": Examining ai/ml practitioners' challenges during co-production of responsible ai values. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3580903. URL https://doi.org/10.1145/3544548.3580903. 5.1

[258] Michael Veale, Max Van Kleek, and Reuben Binns. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *2018 chi conference on human factors in computing systems*, pages 1–14, 2018. 2.1, 4.1, 5.1

[259] Ben Wagner. Ethics as an escape from regulation. from "ethics-washing" to ethics-shopping? 2018. 4.1, 4.6.2

[260] Travis L Wagner and Ashley Blewer. "the word real is no longer real": Deepfakes, gender, and the challenges of ai-altered video. *Open Information Science*, 3(1):32–46, 2019. 2.2.1

[261] Daisuke Wakabayashi and Scott Shane. Google will not renew pentagon contract that upset employees. *The New York Times*, June 2018. Accessed: 2021-07-1. 2.1, 2.4.3

[262] Robert Philip Weber. *Basic content analysis*. Number 49. Sage, 1990. 4.2

[263] Michael Weiss. Profiting from open source. In *15th European Conference on Pattern Languages of Programs*, pages 1–8, 2010. 2.3.1

[264] Robert S Weiss. *Learning from strangers: The art and method of qualitative interview studies*. Simon and Schuster, 1995. 2.2.2, 4.2, 5.2.3

[265] Meredith Whittaker. The steep cost of capture. *Interactions*, 28(6):50–55, 2021. 2.1, 4.6.4

[266] David Gray Widder. Gender and Robots: A Literature Review, June 2022. 6.1.2

[267] David Gray Widder and Dawn Nafus. Dislocated accountabilities in the "ai supply chain": Modularity and developers' notions of responsibility. *Big Data & Society*, 10 (1):20539517231177620, 2023. 3

[268] David Gray Widder, Dawn Nafus, Laura Dabbish, and James Herbsleb. Limits and possibilities for "ethical ai" in open source: A study of deepfakes. In *ACM conference on fairness, accountability, and transparency*, 2022. 2

[269] David Gray Widder, Derrick Zhen, Laura Dabbish, and James Herbsleb. It's about power: What ethical concerns do software engineers have, and what do they (feel they can) do about them? In *Proceedings of the 2023 ACM Conference on Fairness, Accountability,*

*and Transparency*, pages 467–479, 2023. 4

[270] Phil Wilkinson. A brief history of serious games. In *Entertainment Computing and Serious Games: International GI-Dagstuhl Seminar 15283, Dagstuhl Castle, Germany, July 5-10, 2015, Revised Selected Papers*, pages 17–41. Springer, 2016. 5.1

[271] Amy A Winecoff and Elizabeth A Watkins. Artificial concepts of artificial intelligence: Institutional compliance and resistance in ai startups. *arXiv preprint arXiv:2203.01157*, 2022. 3.4.2

[272] Langdon Winner. Do artifacts have politics? *Daedalus*, pages 121–136, 1980. 3.2, 6.3

[273] Rachel Winter and Anastasia Salter. Deepfakes: uncovering hardcore open source on github. *Porn Studies*, 7(4):382–397, 2020. 2.2.1, 2.3.1

[274] Marty J Wolf, Kevin Bowyer, Don Gotterbarn, and Keith Miller. Open source software: intellectual challenges to the status quo. *ACM SIGCSE Bulletin*, 34(1):317–318, 2002. 2.4.2

[275] Marty J Wolf, Keith W Miller, and Frances S Grodzinsky. On the responsibility for uses of downstream software. *Computer Ethics-Philosophical Enquiry (CEPE) Proceedings*, 2019(1):3, 2019. 2.4.3

[276] Richmond Y Wong. Tactics of soft resistance in user experience professionals' values work. *ACM Conference on Human-Computer Interaction*, 5(CSCW2):1–28, 2021. 3.3.2, 4.1, 4.6.1, 4.6.2, 4.6.4, 5.1, 6.2, 6.3

[277] Richmond Y. Wong, Michael A. Madaio, and Nick Merrill. Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics. *ACM Conference on Human-Computer Interaction*, 7(CSCW1):1–27, April 2023. doi: 10.1145/3579621. 1.1, 1.3, 4.1, 4.6.2, 5.1, 5.6, 6.2

[278] Chloe Xiang. Eating Disorder Helpline Disables Chatbot for 'Harmful' Responses After Firing Human Staff. https://www.vice.com/en/article/qjvk97/eating-disorder-helpline-

disables-chatbot-for-harmful-responses-after-firing-human-staff, May 2023. 1.2, 6.4

[279] Karen Yeung, Andrew Howes, and Ganna Pogrebna. Ai governance by human rights-centred design, deliberation and oversight: An end to ethics washing. *The Oxford Handbook of AI Ethics, Oxford University Press (2019)*, 2019. 4.1

[280] Robert K Yin. *Case study research and applications: Design and methods*. Sage publications, 2017. 2.2.1, 6.1.1

[281] Meg Young, Michael Katell, and P.M. Krafft. Confronting Power and Corporate Capture at the FAccT Conference. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1375–1386, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533194. 4.6.4

[282] Shurui Zhou, Bogdan Vasilescu, and Christian Kästner. What the fork: a study of inefficient and efficient forking practices in social coding. In *2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 350–361, 2019. 2.1, 2.3.1, 2.4.3

[283] Jichen Zhu and D Fox Harrell. System intentionality and the artificial intelligence hermeneutic network: the role of intentional vocabulary. In *Proceedings of the 2009 Digital Art and Culture Conference*, 2009. 1.2

[284] Thomas Zimmermann. Card-sorting: From text to themes. In *Perspectives on data science for software engineering*, pages 137–141. Elsevier, 2016. 4.2, 5.2.4

[285] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*, 2017. 1.3

[286] J Zunger. Computer science faces an ethics crisis. the cambridge analytica scandal proves it. *Boston Globe*, 22, 2018. 4.6.2