Power and Play:
Investigating "License to Critique" in Teams' AI Ethics Discussions

DAVID GRAY WIDDER, Digital Life Initiative, Cornell Tech, USA
LAURA DABBISH, Carnegie Mellon University, USA
JAMES HERBSLEB, Carnegie Mellon University, USA
NIKOLAS MARTELARO, Carnegie Mellon University, USA

Past work has sought to design AI ethics interventions–such as checklists or toolkits–to help practitioners design more ethical AI systems. However, other work demonstrates how these interventions may instead serve to *limit* critique to that addressed within the intervention, while rendering broader concerns illegitimate. In this paper, drawing on work examining how standards enact discursive closure and how power relations affect whether and how people raise critique, we recruit three corporate teams, and one activist team, each with prior context working with one another, to play a game designed to trigger broad discussion around AI ethics. We 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 discussions. We then report on how particular affordances of this game may influence discussion, and find that the hypothetical context created in the game is unlikely to be a viable mechanism for real world change. We discuss how power dynamics within a group and notions of "scope" affect whether people may be willing to raise critique in AI ethics discussions, and discuss our finding that games are unlikely to enable direct changes to products or practice, but may be more likely to allow members to find critically-aligned allies for future collective action.

CCS Concepts: • **Social and professional topics** → *Codes of ethics*; • **Software and its engineering** → *Designing software*; • **Human-centered computing** → *Empirical studies in HCI*.

Additional Key Words and Phrases: tech ethics, critique, games, interpersonal power, ethics discussion

ACM Reference Format:
David Gray Widder, Laura Dabbish, James Herbsleb, and Nikolas Martelaro. 2024. Power and Play: Investigating "License to Critique" in Teams' AI Ethics Discussions. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 399 (November 2024), 23 pages. https://doi.org/10.1145/3686938

1 Introduction and Related Work

Many companies have Responsible AI guidelines, which often revolve around principles like Fairness, Accountability, and Transparency [28]. However, research finds that structural and systemic limits encoded in Silicon Valley logics, such as technosolutionism and the normalization of failure, limit those who seek to enact change [42]. High-profile incidents illustrate these limits. In separate incidents, Meredith Whittaker [13] and Timnit Gebru [43] departed Google after seeking to organize against building military drone tech and the company's handling of sexual harassment, and bias and environmental impacts of ever-larger language models, respectively. Such cases demonstrate that there are limits to the kind of direct ethical critique acceptable within tech companies.

Authors' Contact Information: David Gray Widder, Digital Life Initiative, Cornell Tech, New York City, USA, david.g.widder@gmail.com; Laura Dabbish, Carnegie Mellon University, Pittsburgh, USA; James Herbsleb, Carnegie Mellon University, Pittsburgh, USA; Nikolas Martelaro, Carnegie Mellon University, Pittsburgh, USA.

ACM 2573-0142/2024/11-ART399
https://doi.org/10.1145/3686938

The prolific adoption of Responsible AI standards in companies [28] may initially seem to legitimize workers as they raise ethical concerns [36]. However, other work shows that they may, in fact, have the side effect of setting specific boundaries on what is and is not legitimate to raise as an ethical concern. In their work examining environmental sustainability standards in companies, Christensen *et al.* show how such standards enact “discursive closure,” which they understand in light of Deetz [16] to mean legitimizing certain narrow kinds of employee critique while tacitly ruling others out of scope [11]. 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 regarding the effect of responsible AI standards. For example, Keyes *et al.* satirically argue that a system can be Fair, Accountable, and Transparent yet still “turn the elderly into high-nutrient slurry,” showing how certain harms may be outside of the scope of discursively closed principles [31]. Greene *et al.* show that the language of Responsible AI standards focuses scrutiny on the *design* of an AI system, but this focus thereby “reject[s] critiques of business practice” [20]. In this sense, principles such as Fair, Accountable, or Transparent that define responsible AI standards, or their focus on AI design rather than underlying business logics may set the discursive limits of the *de facto* language of AI ethics, thereby rendering concerns outside of this language less legitimate or intelligible,[1] both for workers in companies and for activists and policymakers seeking to influence companies’ actions. Put simply, if an ethical issue is not addressed by a guideline or standard, then it may be seen as out of bounds for critique, thereby limiting which issues can be discussed. In particular, standards and guidelines often seek to help practitioners *build better* AI systems [20], instead of helping them to critique the (business) ends to which they may be put, limiting critique to technical implementation instead of questioning if the system should be built at all.

While past research shows that information sharing among team members predicts outcomes of effectiveness such as solution correctness, profitability or market growth [41], and highlights the complexity of team dynamics through analysis of car manufacturing environments [5], some studies have examined team discussions of AI ethics, a context where it may be less clear what constitutes a “correct” answer, or where different actors may evaluate the ethics of different outcomes differently. For example, prior research has examined how teams implement AI ethics guidelines (*i.e.*, [28]) into organizational processes [45], uncovering structural and systemic limits such as role uncertainty where practitioners are unsure about their remit, and misaligned incentives between teams. To overcome these issues, some have explored how workers can try to steer conversations within their organizations. For example, Madaio *et al.* [36] examine checklists as a way to enable team discussion on designing fair AI systems. Their study found that some workers were concerned that advocating for AI fairness issues outright 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 [36]. Related work on UX professionals seeking to steer their company’s values show how they encounter similar limits, and thus resort to “soft” tactics to attempt to change values and practices while still operating within perceived acceptable bounds within their company [62]. Other structural and systemic limits include insufficient power within one’s role to address ethical concerns [45, 60]. Furthermore, many AI ethics tools or guidelines are often focused on a subset of technical machine learning topics [63]—such examples from CSCW include systems for leveraging crowds to develop ethical constraint specification of AI systems [38] or observing how datasheets for datasets may support ethical thinking [9]. Such systems

[1] Indeed, this constrained language of AI ethics may operate similarly to the *de facto* language of “doing diversity,” in a way that similarly “conceals inequalities and neutralises histories of antagonism and struggle.” [2]

can improve the ethical implementation of AI and workers' abilities to engage with ethical AI considerations, at least within the technical areas such tools aim to address.

However, given the concerns about the discursive closure effects of standards and sets of principles outlined above [11], 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 broader culture in which they are enacted—organizations with certain notions of "efficiency", technological solutionism, and status hierarchies based on technical merit [42]—may limit what concerns workers feel able to raise under these kinds of interventions. Checklists and other artifacts that 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 [63].

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 [24, 36, 53, 55, 59], 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 [35], foreclosing 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. Past ethnographic research examines similarly sensitive questions in public sector and medical contexts, for example, the causes of the Challenger disaster at NASA [54], trust in AI tool use for space mission software at NASA [57], secret nuclear weapons development at a National Laboratory [22], and cultural barriers in efforts to shorten doctors' shifts in hospitals to improve patient safety [30]. This work grapples with and successfully negotiates questions of researcher access in order to enable long-term study of similar questions. However, these may be difficult to transfer to for-profit companies: public sector institutions often have mechanisms for public oversight and scrutiny, and medical contexts have existing professional norms, fiduciary duties, and legal accountability that do not exist in AI companies [44]. 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. Anecdotally, many teams we approached were reluctant to grant us access, citing the sensitive nature of their work.

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). We recruited three teams across two companies and one activist group. Recruiting real teams allows them to bring their associated shared experiences and context, shared understanding of process, and existing power dynamics into the study discussions. Including an activist group provides a point of comparison from which to question norms in company contexts, so as to imagine other possible practices. 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 discussions in their ordinary team context. In short, we set out 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 sought 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” [46]. 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” [46].

Motivated by Scott’s concept of offstage talk [46], we see a connection with games based around speculative futures as a way to provide an “innocuous” context for discussion. In this work, we explore how a game can be used as a method to study conversation around AI ethics and to shape conversations between team members.

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” [17]. 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 [39]. In a technology context specifically, *Project Amelia* used immersive theater to encourage participants to reflect on their privacy behavior within technology [47]. Past research reviews and draws together Speculative Design with Games Studies approaches, concluding that “explorative worlds produced by games [have] much to offer over the traditional mediums of speculative design due to their inherent interactivity” [14], and a past meta-analysis demonstrates how games have been used as data collection methods [48]. Flanagan and Nissenbaum have also shown how digital games embed values, and may be used in “animating personal, political, and artistic expression” [18].

There are also at least two existing examples of past works proposing games which employ speculative approaches in an AI ethics context. Ballard *et al.*’s *Judgment Call* was designed around Microsoft’s articulated ethical principles to create “space for difficult or uncomfortable conversations” [3] and argued for the applicability of design fiction methods in game contexts. 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 emerging from these new combinations. To examine the possibility for these kinds of games to create the “innocuous” understandings that Scott wrote of [46], and to examine how they may work against discursive closure [11], 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 due to existing power dynamics and the disconnect between low-stakes hypotheticals and higher stakes changes to the product. However, we also see that games may help teammates better understand each others’ 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 [61]. Our study contributes a critical evaluation of the limits and opportunities for using games as a tool to help professional teams discuss sensitive issues like AI ethics (RQ2), and uses this tool as a probe in data collection to investigate factors that influence team members' "license to critique" when discussing such issues in their ordinary work (RQ1).

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.

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 & Ju'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 [3], but we choose this one because its source materials are readily available, can be easily adapted by adding cards, includes an online version for remote participants,[2] 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 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.

In our study, gameplay sessions lasted 1.5 hours. Participants played through many rounds with different *Prompts*. There was no time limit set for each round, with groups completing between three to eight game rounds each.

2.2 Participants

We recruited 16 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[3], and the activist group played in person using a printed deck, reflecting how each of these teams ordinarily met together. We note that past work shows that virtual versus in-person

[2] https://github.com/nikmart/what-could-go-wrong-ai
[3] www.playingcards.io

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 1. Participant demographics.

environments affect how teams discuss creative ideas [10], and we discuss how affordances from virtual environments affected discussions in Sections 3.1 and 4.2.1.

To protect anonymity and reduce re-identification risk, some demographic descriptors here have been generalized in Table 1, and we provide group rather than individual identifiers alongside quotes, even where quotes originated from individual follow-up interviews: *i.e.*, C2-T1 refers to a quote from a participant at Company 2 Team 1, and AC refers to a participant in the Activist Collective. Where aspects of a participant's identity are important for contextualizing what they say (*i.e.*, how one's gender affects their ability to raise critique), we selectively note this context adjacent to their quotes, thereby attempting to balance context with anonymity. While team C1-T2 played with their manager (which we discuss in Section 3.3.1), no other company teams did, but we note significant diversity in team seniority (which we discuss in Section 3.3.3) as can be seen in Table 1.

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 [56] recorded follow-up interview (lasting an average of 34 minutes, but as long as 49 minutes or as short as 22 minutes). During the interview, we asked participants to reflect on in-game experiences and conversations which arose, but with a special focus on contrasts between in-game discussions and past AI-ethics discussions within their team or organizational context. We also focused on the extent to which they did or did not feel comfortable raising anything or disagreeing, during in-game or prior AI ethics discussions. For remote sessions, researchers turned off their cameras unless answering questions to minimize any effect their presence may have had on the teams' discussion.

2.4 Data analysis

We analyzed data under an interpretive epistemological paradigm [34] using thematic analysis [12]. The first author began by open coding [49] 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 [12] comprising 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 [65]. The nature of our data collection and interpretive epistemological stance yields rich insight into participants' experiences and how they make sense of them, but results should not be directly generalized into new contexts without further study.

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

In this section, we discuss factors that influence team members' license to critique when discussing AI ethics. To set the stage, we first show how techno-optimistic organizational norms mean that critique is often perceived as "too negative", after which we narrow in on two particular ways that this occurs. Firstly, license to critique is regulated through a notion of "scope", bounding the issues a team believes it can or will take on, enacted pragmatically through time pressures which curtail discussion of "edge case[s]" and to issues solvable through "technical solution[s]", reified by role divisions which compartmentalize ethical concerns. Secondly, we show how "who is in the room?" matters for resulting ethics discussions, and is affected by whether one's manager is present, one's understanding of teammates' orientation toward critical topics, and the ways that discussants' personal identity (*e.g.*, age, gender, experience) affects who is credibly able to raise concerns. In this section we rely predominantly on data collected during one-on-one follow-up interviews, with participants' experiences during the game secondarily serving as a probe to enable reflection and contrasts to ordinary discussions of ethics within their organizations.

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" (C1-T2)* before she would *"have a conversation about it" (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 and techno-optimism surrounding new technologies. 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" (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" (C2-T1).*

Another participant who identified as a woman 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" (C1-T1)*, but that she was recently *"trying to be a 'team player', and really trying to hold back when I disagree" (C1-T1).* After being told by a close confidant that *"you complain a lot [...] you shouldn't do that" (C1-T1)* she has *"been trying to ramp back on disagreements and save it for when I feel most passionately" (C1-T1).* Her particular concern about being told she is "too negative at work" alludes to past work examining how women are viewed as themselves a problem when they speak up about problems at work [1]. We further discuss relationships between identity and license to critique in Sec. 3.3.3.

3.1.1 Summary. In this opening subsection we have seen how, broadly, organizational norms enact discursive closure [11] whether through an implicitly understood need for concerns to be "significant" to merit discussion, as well as more quotidian worries about frustrating excited colleagues, or being perceived as too negative.

3.2 "Scope", its contestations, and its effects

From our interviews, we saw the notion of *scope* invoked to set boundaries on what corporate teams believe they can or will take action on, and thus where discussion is focused. Scope 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.

3.2.1 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" (C2-T1)*, often focused on specific aspects or features of a product. Another 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" (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" (C1-T2)*, but *"those are things that I probably wouldn't share in an Ethical Impact Assessment review" (C1-T2)*, because it wouldn't be "applicable."

Participants perceived that certain kinds of harms or remedies, such as climate harms or those requiring "diffuse" systemic change 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."' (C2-T1)* Another suggested that *"AI ethics is that it often falls short of trying to work towards systemic change" (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]" (AC).*

However, some ethical issues were seen as so potentially damaging to the product success, that they ceased 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]" (C2-T1).* This suggests that something being an "ethical question" may be its own kind of limiting or minimizing scope.

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

3.2.3 *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 [11], participants discussed 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, such technosolutionist or technochauvinist thinking can entrench existing inequities, and obscure other ways of thinking or conceiving of problems or solutions [15]. This demonstrates how organizational dynamics and structure limits the kind of solutions seen as possible, by fracturing questions of accountability[58] between "what to build" questions and the narrower and technical scope of "how to build" that teams felt power over. 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" (C2-T1)*, comparing this to in-game discussion which *"felt more generative [because] there wasn't as much of a goal" (C2-T1).* A participant also noticed how conversations often *"jump towards like, what's, what's a good technical solution?" (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" (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 closure: *"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' [they] tend to be a lot more directed" (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" (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” (C1-T2)*

Participants suggested a variety of possible reasons for this rush to discuss technical solutions. One suggested that in a *“tech company” (C2-T1)*, with an *“engineering mindset” (C2-T1)*, open-ended introspection is *“not always the vibe” (C2-T1).* One participant reasoned that this may be due to *“what is possible to change” (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” (C2-T1).*

3.2.4 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[s] or address[es] those topics on [our] projects [...] we have people that do that” (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” (C1-T1).* Ethics was seen as a function of one narrow specialist committee, and called out as an example of a broader phenomenon of the compartmentalization of horizontal capabilities, thus demonstrating how organizational dynamics narrow both who feels able to raise ethical critique, and how broadly or narrowly segmented this critique is able to be.

Participants spoke about countering this compartmentalization, saying they *“we need more diverse thoughts here” (C1-T1)* and seek to *“strengthen the bonds among some of the product, ML product practitioners, and me [in her ethics-focused role]” (C2-T1).* In this way, we see different ways that participants construct their role and and professional duties with respect to ethics—either as something that is best left to teammates assigned specifically to do this work, or conversely, as something that should not be compartmentalized, thereby still recognizing compartmentalization as a default practice bounding their own role, and hoping to work against this.

3.2.5 Summary. In this subsection we have shown how an often implicitly understood notion of “scope” mediates what kind of critique is rendered acceptable, thereby constituting that which is rhetorically “closed” in the enactment of discursive closure [11]. We have further shown how scope is enacted and maintained in organizational and work practices, including how *time pressures* lead to discursive closure due to perceived relevance, a *push to solve* issues focusing discussion on issues addressable within technical changes in the project at hand, and how *role divisions* lead to a compartmentalization effect where some see ethics questions as not within the scope of their role.

3.3 “Who is in the room”: participants’ power and critical orientation

3.3.1 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 of 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” (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” (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” (C1-T1).* One participant commented on the *“surveillance technology that’s on everyone’s [company] laptops” (C2-T1)*, and also on *“worker exploitation” (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” (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.'" (C2-T1)*. Others reported their *"manager, and like, my skip level [managers]" (C2-T1)* encouraging her to prioritize work that *"they felt would be more impactful in a product" (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" (C1-T2)*, and another said that given their past experience working in a formal ethics team, *"we're all peers [...] I knew I could share freely in front of this group of people" (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, bringing forth perhaps the most direct example of how particular team dynamics affect how critique can be raised.

3.3.2 Awareness of teammates' critical orientation. Participants displayed awareness of who was in the room, and their ethical views and critical orientation. We define *critical orientation* to mean an understanding of a teammate's (un)willingness to engage in ethical critique that may challenge existing project or business goals, as which may be risky [36, 60]. We saw that participants' awareness of others teammates' critical orientation affected how they expressed themselves or whether they raised 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" (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" (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" (C1-T1)*, because they wouldn't have a *"shared baseline" (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" (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" (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" (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" (AC)*.

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" (AC)*.

3.3.3 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" (C1-T1)* because *"he is really young. And [...] he joined very recently" (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" (C1-T1). Another participant whose play session included only female-presenting 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" (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 woman-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" (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 questioned whether 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" (AC).* Another participant felt they might not qualify to participate in this study *"one of the requirements in [study criteria] was to be like in an engineering field. So I was like, I'm not that" (AC).* This theme relates both familiar team dynamics, in the way that particular identities (*i.e.*, gender, age) were looked on with credibility, but also organizational dynamics more salient to tech companies where those with engineering expertise are seen as most credible.

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

3.3.4 Summary. In this subsection we have shown the various ways that the particularities of who is in the room affects one's license to critique [11] in discussions of AI ethics. This includes power dynamics resulting from one's manager being present, awareness of the critical orientation of colleagues in the room, and how personal attributes like age, gender, and expertise affect perceptions of who is able to raise critique.

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

In this section, we examine how affordances from the game context appeared to affect how conversations unfolded. Following from our earlier discussion on how notions of scope limit license to critique (Sec. 3.2), we firstly show how the introduction of randomness, discussion about harms beyond a particular system, and the creation of a hypothetical context allowed a wider scope of discussion. However, we also show that participants felt that this expanded scope in a hypothetical context would not transfer back to discussions of actual projects, which are treated as compliance exercises and risk painting one's project in a negative light. Secondly, we demonstrate how the game enabled teammates to learn more about each other by creating a space to socialize, to be vulnerable, to learn about other's critical orientation, and to find "allies" for future discussions of ethics back in one's real work environment. Here, we rely primarily on observations and recordings of the game session. We examine how the game was able to expand scope, how participants remixed rules, and how they used the game context as an opportunity to learn about teammates' past experiences and critical orientation.

4.1 Expanding scope

4.1.1 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" (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" (C1-T2).* Another mentioned that a format where *"there's not really a correct answer" (C2-T1)* is one she hadn't considered before, but noted that it *"got a lot of us thinking in different directions" (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" (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 [...] that someone's working on something [concerning]. They're not necessarily, like, speculative" (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" (AC),* and people appeared to feel able to speak from this frame of reference throughout game play.

4.1.2 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? " (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" (AC).* 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" (AC).*

4.1.3 Hypothetical situations as an "innocuous" context for discussion. Several participants brought up how the hypothetical context of the game provided an "innocuous" context [46] to raise critique that they felt may have otherwise been 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" (C1-T2)* then *"in* *that context, it probably does make it easier" (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!'" (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'" (C1-T2)*, that when conversation is *"taken away from the project that we were doing [...] everyone was very free" (C1-T1).*

Others corroborated this, suggesting that raising concerns *"hypothetically in a game [...] is really nice because it's just 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" (C1-T2).* One said the point is to "make assumptions": *"I saw this one [card] is, you know [...] I'm making a lot of assumptions here. But that I think that's maybe the point of some of these discussions" (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'" (C1-T2).*

4.1.4 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 [...] people start to become more defensive. They don't expand into all the things that can go wrong" (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 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" (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" (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" (C2-T1)*, which appeared to set the expectation that this is not the context from which immediate or actionable changes to the product or process in work contexts would arise.

4.2 Learning about teammates

4.2.1 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[ful], like it was, it was really shocking. Like 'Oh, wow' like he was part of it. Like he was there" (C1-T1).* Another participant suggested that with an *"all-audio [meeting] culture"* 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" (C1-T1).* This speaks how perhaps team dynamics within companies, by default, do not enable vulnerability nor space to talk beyond work tasks, both of which appeared to be helpful when discussing ethics.

4.2.2 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 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 fight" (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" (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 [... and because of this,] theory of mind can be difficult to achieve" (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" (AC).*

We note that exposing one's critical orientation, especially around those one does not know well, may be a risky endeavor: one risks being labeled a troublemaker [1] or facing career repercussions [36]. Team dynamics which better enable one to feel safe exposing views that may challenge business logics may alleviate the sense of isolation that past work shows people can feel when holding such views [60], or the "malaise" that may "permeate" through organizations when this is more widespread [50] We return to this further in Section 5.1, below.

5 Discussion

Our results show that in the 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 or colleagues of unknown critical orientation are in the room, 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 [46]. 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 [8, 19, 29, 33, 60] which makes clear that research on AI Ethics must be more attentive to the differentiated power relationships between those that may use, request, or engage in proposed AI ethics interventions or discuss AI ethics issues, and how these power relationships are ingrained in organizational conventions and culture. If future empirical work examining how those discussing AI ethics in any context does not attend to power in their analysis, such work risks missing major determinants of any apparent agreements or disagreements that may arise.

5.1 A hypothetical game context is unlikely to lead to direct change, but may help find critically-aligned allies

We join past work in discussing how tech industry logics such as technological solutionism affect ethics initiatives [42], and other topics intersecting with ethics. These include privacy, where some worry that "advocating for privacy might be indirectly misaligned with career incentives" [32]; accessibility, where this is not seen as a first-order concern and only attended to if customers ask for it [60], or studies which show that implementing accessibility is an individual rather than an institutional responsibility [6]; and sustainability, where the "endless pursuit of achieving higher model quality has led to the exponential scaling of AI with significant energy and environmental footprint implications" [64], especially in the context of increasingly Large Language Models [4, 21].

Our work casts doubt on whether an innocuous context, such as those created by games, may enable discussions that successfully challenge these logics and transfer to changes in a team's real-world context. In short, our findings suggest that games are unlikely to present opportunities to escape the power dynamics that shape AI ethics discussion in ways that lead directly to project changes. 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" [46], raising questions about whether game-based AI Ethics interventions may expand the scope of what is sayable "on-stage" and create such contexts to make dissent safer, our results tell a more complicated story. As we detail in Section 4.1.3, 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 the utility of games and other hypothetical interventions to create space for discussions or agreements that 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 suggest that interventions such as games may be unlikely to meaningfully intervene in power dynamics in ways that directly spur real-world action, given organizational logics and policies. 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" [40], 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" [3]. 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 and 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 (see Sec. 3.3.2). 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" [46] 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 [46], and we join Scott to suggest that collective accounts are where solidarity begins. It is nonetheless important to note that games are, at best, a modest intervention, perhaps providing a possible context for solidarity to arise, among scant alternative spaces.

While formal AI ethics activities such as checklists may be able to "empower *individual* advocates" (emphasis added) by legitimizing a particular issue [36] 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 [11]. Past studies demonstrate 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" [60], and our results suggest that "innocuous" contexts (*i.e.*, [46]) 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" [27]. While the literature on this construct often focuses on the impact of cross-understanding for narrower questions of "product quality" and avoids questions that members would find ”technically, politically, or otherwise unacceptable" [25], parallels may be drawn beyond quality and features, to questions of product ethics.

Feminist theory helps illuminate the distinction between hypothetical or "innocuous" [46] contexts enabling real-world changes directly versus enabling stronger ties, which may then become a 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 [23]. Extending Haraway's 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" [51]. Rather than theorizing games as separate safe spaces from which to speak from nowhere, we suggest that games may be a modest but deliberately created opportunity for different partial knowledges to meet, learn what they have in common, and enable "collective knowledge of the specific locations of our respective visions" [51], from which durable coalitions and collectivities for action may arise. Drawing on both Haraway and Suchman, Widder and Nafus show how social ties, responsibilities, and concerns—developed outside of engineers' assigned duties—are the basis for the (little) AI ethics work that does get done, and innocuous contexts created by games may enable non-work contexts for these ties to form and strengthen [58].

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.

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

Our results show how notions of "scope"—what a team believes it can or will take action on—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 3.2 and 4.1). Firstly, we show 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 Widder and Nafus argue [58], but demonstrates 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, 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: being presented with particular systems they are supposed to evaluate for ethical issues, and propose “mitigations” to solve said issues, and this frame makes it harder to discuss how the systems they’re evaluating relate to one another, or how they may require “systemic changes” that this frame 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 (see Section 4.1), and how people may employ rhetorical moves to frame ethics questions as larger scoped product questions.

We theorize scope as a softer way to dismiss critique. This functioned as an instance of problem closure [26], 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 dismisses those who may wish to keep it open or believe it to be unsolved. Our data shows that 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 the need to deny the validity of an issue outright but instead suggesting it is out of scope, one can avoid 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” [52]. Instead of arguing that an issue is false, arguments based on jurisdictional stasis question the “jurisdictional appropriateness of the issue” [52], 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.* [11] 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”’ [20]. 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” [31].

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” [37], 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.

6 Conclusion

Past work has sought to design AI ethics interventions–such as checklists [36] or toolkits [7]–to help practitioners design more ethical AI systems. However, other work demonstrates how these interventions [63] and the principles they're based on [20] may serve to instead limit critique to those addressed within the intervention, while rendering broader concerns illegitimate, and how core logics of the tech industry make raising ethical concerns that reach beyond technological solutionism or market fundamentalism [42] to challenge business practice [60] extraordinarily difficult.

In this paper, we examined how teams discuss AI ethics issues by drawing on past work examining how standards in other contexts enact discursive closure [11], and on how power relations affect whether and how critique is raised [46]. We recruit three corporate teams, and one activist team, each with prior context with one another, 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 organizational and team dynamics 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. Our finding suggest 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.

7 Acknowledgments

We are grateful to Wendy Ju, who helped develop the *What Could Go Wrong?* card game with Nikolas Martelaro. We thank Justin Deyo, Katie Ciez and Grace Klein at the Carnegie Mellon University Block Center for Technology and Policy for hosting early pilot tests, and to student research assistants Sherry Chen, Eunbea Ji and Wan Tong Li for helping run these tests. We are grateful to our anonymous participants, and to the three anonymous reviewers and Michael Madaio, Dawn Nafus, Lucy Suchman, and Julie Martinson Widder, who gave feedback on earlier drafts, as well as the support of the Digital Life Initiative at Cornell Tech. This material is based upon work supported by the Software Engineering Institute at Carnegie Mellon University, the National Science Foundation under Grant No. #2107298, and the Alfred P. Sloan Foundation under award G202014036.

References

[1] Sara Ahmed. 2021. *Complaint!* Duke University Press.

[2] Sara Ahmed and Elaine Swan. 2006. Doing Diversity. *Policy Futures in Education* 4, 2 (June 2006), 96–100. https://doi.org/10.2304/pfie.2006.4.2.96

[3] Stephanie Ballard, Karen M. Chappell, and Kristen Kennedy. 2019. 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*. ACM, San Diego CA USA, 421–433. https://doi.org/10.1145/3322276.3323697

[4] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623.

[5] Jos Benders and Geert Van Hootegem. 1999. Teams and Their Context: Moving the Team Discussion Beyond Existing Dichotomies. *Journal of Management Studies* 36, 5 (1999), 609–628. https://doi.org/10.1111/1467-6486.00151

[6] Tingting Bi, Xin Xia, David Lo, John Grundy, Thomas Zimmermann, and Denae Ford. 2022. Accessibility in Software Practice: A Practitioner's Perspective. *ACM Transactions on Software Engineering and Methodology* 31, 4 (Oct. 2022), 1–26. https://doi.org/10.1145/3503508

[7] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).

[8] Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. 2022. The Forgotten Margins of AI Ethics. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 948–958. https://doi.org/10.1145/3531146.3533157

[9] Karen L. Boyd. 2021. Datasheets for Datasets Help ML Engineers Notice and Understand Ethical Issues in Training Data. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 438 (oct 2021), 27 pages. https://doi.org/10.1145/3479582

[10] Melanie S. Brucks and Jonathan Levav. 2022. Virtual Communication Curbs Creative Idea Generation. *Nature* 605, 7908 (May 2022), 108–112. https://doi.org/10.1038/s41586-022-04643-y

[11] Lars Thøger Christensen, Mette Morsing, and Ole Thyssen. 2017. License to Critique: A Communication Perspective on Sustainability Standards. *Business Ethics Quarterly* 27, 2 (April 2017), 239–262. https://doi.org/10.1017/beq.2016.66

[12] Victoria Clarke and Virginia Braun. 2021. Thematic analysis: a practical guide. *Thematic Analysis* (2021), 1–100.

[13] Kate Conger and Daisuke Wakabayashi. 2019. Google Employees Say They Faced Retaliation After Organizing Walkout. https://www.nytimes.com/2019/04/22/technology/google-walkout-employees-retaliation.html. *The New York Times* (22 April 2019). Accessed: 2023-06-27.

[14] Paul Coulton, Dan Burnett, and Adrian Gradinar. 2016. Games as Speculative Design: Allowing Players to Consider Alternate Presents and Plausible Features. In *Design Research Society Conference 2016*. https://doi.org/10.21606/drs.2016.15

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

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

[17] Anthony Dunne and Fiona Raby. 2013. *Speculative Everything: Design, Fiction, and Social Dreaming*. MIT Press. Google-Books-ID: 9gQyAgAAQBAJ.

[18] Mary Flanagan and Helen Nissenbaum. 2014. *Values at Play in Digital Games*. https://doi.org/10.7551/mitpress/9016.001.0001

[19] Ben Gansky and Sean McDonald. 2022. CounterFAccTual: How FAccT undermines its organizing principles. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1982–1992.

[20] Daniel Greene, Anna Lauren Hoffmann, and Luke Stark. 2019. 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*.

[21] Sireesh Gururaja, Amanda Bertsch, Clara Na, David Gray Widder, and Emma Strubell. 2023. To Build Our Future, We Must Know Our Past: Contextualizing Paradigm Shifts in Natural Language Processing. *arXiv preprint arXiv:2310.07715* (2023).

[22] Hugh Gusterson. 1996. *Nuclear Rites: A Weapons Laboratory at the End of the Cold War*. University of California Press.

[23] Donna J Haraway. 1991. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Simians, cyborgs, and women: The reinvention of nature* (1991), 183–201.

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

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

[26] Lee Humphreys. 2005. Reframing Social Groups, Closure, and Stabilization in the Social Construction of Technology. *Social Epistemology* 19, 2-3 (Jan. 2005), 231–253. https://doi.org/10.1080/02691720500145449 Publisher: Routledge _eprint: https://doi.org/10.1080/02691720500145449.

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

[28] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[29] Khari Johnson. 2019. AI ethics is all about power. *Venture Beat* 1 (2019).

[30] Katherine C. Kellogg. 2009. Operating Room: Relational Spaces and Microinstitutional Change in Surgery. *Amer. J. Sociology* 115, 3 (Nov. 2009), 657–711. https://doi.org/10.1086/603535

[31] Os Keyes, Jevan Hutson, and Meredith Durbin. 2019. 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*. 1–11.

[32] Hao-Ping Hank Lee, Lan Gao, Stephanie Yang, Jodi Forlizzi, and Sauvik Das. [n. d.]. "I Don't Know If We're Doing Good. I Don't Know If We're Doing Bad": Investigating How Practitioners Scope, Motivate, and Conduct Privacy Work When Developing AI Products. ([n. d.]).

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

[34] Yvonna S Lincoln, Susan A Lynham, and Egon G Guba. 2011. Paradigmatic controversies, contradictions, and emerging confluences, revisited. *The Sage handbook of qualitative research* 4 (2011), 97–128.

[35] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *ACM Conference on Human-Computer Interaction* 6, CSCW1 (2022), 1–26.

[36] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[37] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Microsoft AI Fairness Checklist. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4t6dA.

[38] Travis Mandel, Jahnu Best, Randall H. Tanaka, Hiram Temple, Chansen Haili, Sebastian J. Carter, Kayla Schlechtinger, and Roy Szeto. 2020. Using the Crowd to Prevent Harmful AI Behavior. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 97 (oct 2020), 25 pages. https://doi.org/10.1145/3415168

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

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

[41] Jessica R. Mesmer-Magnus and Leslie A. DeChurch. 2009. Information Sharing and Team Performance: A Meta-Analysis. *Journal of Applied Psychology* 94, 2 (2009), 535–546. https://doi.org/10.1037/a0013773

[42] Jacob Metcalf, Emanuel Moss, et al. 2019. Owning ethics: Corporate logics, silicon valley, and the institutionalization of ethics. *Social Research: An International Quarterly* 86, 2 (2019), 449–476.

[43] Cade Metz and Daisuke Wakabayashi. 2020. Google Researcher Says She Was Fired Over Paper Highlighting Bias in A.I. *The New York Times* (Dec. 2020).

[44] Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1, 11 (2019), 501–507.

[45] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2020. 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).

[46] James C Scott. 1990. *Domination and the arts of resistance: Hidden transcripts.* Yale university press.

[47] Michael Skirpan, Maggie Oates, Daragh Byrne, Robert Cunningham, and Lorrie Faith Cranor. 2022. Is a privacy crisis experienced, a privacy crisis avoided? *Commun. ACM* 65, 3 (March 2022), 26–29. https://doi.org/10.1145/3512325

[48] Shamus P. Smith, Karen Blackmore, and Keith Nesbitt. 2015. A Meta-Analysis of Data Collection in Serious Games Research. In *Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement*, Christian Sebastian Loh, Yanyan Sheng, and Dirk Ifenthaler (Eds.). Springer International Publishing, Cham, 31–55. https://doi.org/10.1007/978-3-319-05834-4_2

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

[50] Norman Makoto Su, Amanda Lazar, and Lilly Irani. 2021. Critical Affects: Tech Work Emotions Amidst the Techlash. *ACM Conference on Human-Computer Interaction* 5, CSCW1 (2021), 1–27.

[51] Lucy Suchman. 2002. Located accountabilities in technology production. *Scandinavian journal of information systems* 14, 2 (2002), 7.

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

[53] Rama Adithya Varanasi and Nitesh Goyal. 2023. "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* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 251, 17 pages. https://doi.org/10.1145/3544548.3580903

[54] Diane Vaughan. 1996. *The Challenger launch decision: Risky technology, culture, and deviance at NASA.* University of Chicago press.

[55] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. 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*. 1–14.

[56] Robert S Weiss. 1995. *Learning from strangers: The art and method of qualitative interview studies.* Simon and Schuster.

[57] David Gray Widder, Laura Dabbish, James D Herbsleb, Alexandra Holloway, and Scott Davidoff. 2021. Trust in Collaborative Automation in High Stakes Software Engineering Work: A Case Study at NASA. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[58] David Gray Widder and Dawn Nafus. 2023. Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big Data and Society* (2023), 1–12. https://doi.org/10.1177/20539517231177620

[59] David Gray Widder, Dawn Nafus, Laura Dabbish, and James Herbsleb. 2022. Limits and Possibilities for "Ethical AI" in Open Source: A Study of Deepfakes. In *conference on fairness, accountability, and transparency*.

[60] David Gray Widder, Derrick Zhen, Laura Dabbish, and James Herbsleb. 2023. 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 (ACM FAccT)*. ACM, Chicago IL, USA. https://doi.org/10.1145/3593013.3594012

[61] Phil Wilkinson. 2016. 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*. Springer, 17–41.

[62] Richmond Y. Wong. 2021. Tactics of Soft Resistance in User Experience Professionals' Values Work. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 355 (oct 2021), 28 pages. https://doi.org/10.1145/3479499

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

[64] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga Behram, James Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin S. Lee, Bugra Akyildiz, Maximilian Balandat, Joe Spisak, Ravi Jain, Mike Rabbat, and Kim Hazelwood. 2022. Sustainable AI: Environmental Implications, Challenges and Opportunities. arXiv:2111.00364 [cs]

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

Received July 2023; revised January 2024; accepted March 2024