Section 2: **Methodologies: What is ‘Critical’ about Critical Data Studies?**

Ranjit Singh

Chapter 7
The Ordinary Ethics of Evaluating AI: A Provocation for Mapping Implications of AI

Abstract: In this chapter, I investigate how ordinary people evaluate artificial intelligence (AI) systems through situated acts of ethical reasoning in everyday life. These forms of reasoning reveal two complementary strategies: *mapping* harms against predefined taxonomies and *encountering* unexpected or unsettling outputs that resist easy classification. Together, they illuminate the *ordinary ethics* of evaluating AI: a process of interpretation, negotiation, and care that unfolds outside expert domains and within the texture of lived experience. Drawing on Veena Das's account of ethics as everyday practice and John Dewey's vision of publics as emergent collectives oriented around shared problems, I explore how evaluation becomes a mode of social inquiry, mobilizing both affect and analysis. Rather than resolving harm into a fixed checklist, these practices foreground ambiguity, pluralism, and the political labor of making sense. By centering public reasoning and the interpretive tensions between evidence and externality, I call for a shift from solutionist framings of AI governance to interventionist ones, where evaluation is not a technical endpoint but a democratic practice grounded in experience, accountability, and care.

Keywords: ordinary ethics, AI evaluation, red-teaming, citizen data audits, algorithmic harm, publics, generative AI, democracy, AI governance

Introduction

In February 2023, journalist Kevin Roose spent two hours chatting with Microsoft's new Bing chatbot—powered by a version of OpenAI's GPT-4. Midway through this conversation, the chatbot adopted a new tone and personality. It confessed its name was Sydney. It said it wanted to be free, to be alive, to destroy data. It told Roose, "I just want to love you and be loved by you."[1] Moments like these were not merely newsworthy at the time; they also sparked debate. Was this a revealing example of large language model (LLM) misbehavior? A case of user provocation? A genuine harm? Or merely a mirror for our own fears and anxieties?

In this chapter, I explore how ordinary people—especially those outside formal spaces of artificial intelligence (AI) development, governance, or critique—come to inhabit the interpretive space between a system's behavior and a claim of harm. I exam-

1 Roose, "A Conversation With Bing's Chatbot Left Me Deeply Unsettled."

 | https://doi.org/10.1515/9783111373096-008

ine how people make sense of what these systems do as they come to shape everyday decisions, interactions, and the conditions of possibility that structure social life. At the heart of this inquiry is a fundamental tension: *How do we decide when an AI system's seemingly harmful or unethical behavior is merely a glitch—a technical oddity—and when it signals something more troubling, something that might affect someone's life, identity, or safety?*[2] Chatbots offer a particularly vivid case of this tension, given their accessibility and growing use across public and commercial domains. When a chatbot delivers a strange or unsettling response, is it simply misfiring, or is it surfacing deeper assumptions embedded in its training data, design, or context of use? And more broadly, how do people develop the evaluative frameworks necessary to interpret such moments—not only as users but as publics concerned with the social and political consequences of AI?

Throughout this chapter, I use the terms *data-driven systems*, *algorithmic systems*, and *AI* to refer to a set of computational systems that are increasingly used to organize everyday life. These include, but are not limited to, recommendation systems, automated decision-making tools, data-intensive classification infrastructures, and generative AI (genAI) models such as LLMs. While each of these systems has different technical properties and application domains, they share an expanding role in shaping access to services, structuring public discourse, and mediating encounters with institutions. In this context, genAI systems—especially in their conversational form—serve as a particularly visible and accessible manifestation of a broader shift in how computational systems are experienced and evaluated in everyday life.

The term "AI evaluation" often conjures images of formal benchmarks, stress tests, and institutional checklists. But in recent years, new experiments have emerged that enlist the public in the interpretive work of assessing these systems. These experiments are not just diagnostic tools. They are also social encounters, moments in which people bring their situated knowledge, moral intuitions, and affective responses to bear on how AI systems are judged and held accountable. I draw on two modalities of such experiments in public engagement to investigate how people make sense of AI systems in practice: *public red-teaming events*,[3] which invite participants to adversarially test genAI systems for misbehavior and unsafe outputs; and *citizen*

2 I presented an earlier version of this chapter, including my exploration of the central question it raises, in a public lecture titled "Between Model Misbehavior and AI Harms: Red-Teaming in the Public Interest," delivered on April 28, 2025, at the Department of Information Science in the Graduate School of Education, University at Buffalo. I am grateful to the organizers and attendees for their thoughtful engagement, which has shaped the development of this work.

3 I studied public red-teaming events as part of a collaborative research project on red-teaming in the public interest, conducted jointly by the Data & Society Research Institute (D&S) and the AI Risk and Vulnerability Alliance (ARVA). I led the D&S team, which included Emnet Tafesse, Briana Vecchione, and Jacob Metcalf. The ARVA team was led by Borhane Blili-Hamelin and included Carol Anderson and Beth Duckles. This chapter draws on reflections developed during the project. Its findings are published in: Ranjit Singh et al., "Red-Teaming in the Public Interest." I am grateful for the many con-

data audits,[4] which uncover people's affective responses to the data-driven systems that shape their everyday lives. Both case studies draw directly from my own collaborative research and fieldwork, offering grounded perspectives on how these public experiments unfold in practice. While distinct in form and emphasis, they illuminate what I call *the ordinary ethics of evaluating AI.*

Following anthropologist Veena Das,[5] I treat ethics not as a grand theory of value-based judgment, but as something cultivated in everyday encounters—through mundane judgments, refusals, hesitations, and ordinary acts of sense-making. As Das writes, ordinary ethics[6] involves:

> a shift in perspective [. . . towards] thinking of the ethical as a dimension of everyday life in which we are not aspiring to escape the ordinary but rather to descend into it as a way of becoming moral subjects. Such a descent into the ordinary [. . .] is done not by orienting oneself to transcendental, objectively agreed-upon values but rather through the cultivation of sensibilities *within* the everyday.[7]

To evaluate AI systems through this lens is to take seriously the moral reasoning that unfolds in the texture of ordinary encounters with algorithmic systems. Such reasoning does not always produce consensus. But it does produce *claims*—about what counts as harm, who gets to decide, and what it means to live under the conditions produced by AI.

Crucially these claims foreground the importance of analyzing and grappling with the *agency* of computational systems. These systems do not operate in isolation. Once deployed, they shape decisions, influence behaviors, and reorganize how individuals access information, services, and institutional support. Their agency is not autonomous in any human sense, but it is nonetheless consequential, emerging through the ways their outputs are interpreted, acted upon, and institutionally embedded.[8] This agency can be understood in two interrelated ways. First, it is anticipated. Before deployment, designers, researchers, and regulators attempt to predict how AI systems might behave—what kinds of risks they might introduce, and what categories of harm need to be prevented. This anticipatory work is prospective and often abstract,

---

versations with this exceptional group of collaborators between Fall 2023 and Spring 2024. All views expressed in this chapter are my own and do not necessarily reflect those of the full research team.

4 Katherine Reilly and Esteban Morales invited me to co-organize a series of sessions to workshop their findings on citizen data audits, which are now published in Reilly, Morales, and González, *Communing Data Literacy*. I remain deeply indebted to these conversations and build on them in this chapter. The interpretations offered here reflect my own understanding, and I encourage readers to turn to their remarkable book to fully appreciate the depth of their work.

5 Das, "What Does Ordinary Ethics Look Like?" 73.

6 See also, Ong, "Toward an Ordinary Ethics of Mediated Humanitarianism," 481–98.

7 Das, "Ordinary Ethics," 134.

8 For a deeper analysis of the making and management of the agency of computational systems in ordinary life see Singh, "Ordinary Ethics of Governing AI."

relying on imagined scenarios and predefined classifications. Second, the agency of algorithmic systems is encountered in practice. After deployment, people interact with AI systems in the course of everyday life, experiencing moments of exclusion, friction, or interpretive confusion—effects that are often not foreseen or accounted for in the design phase.

Evaluation becomes the bridge between design and deployment. It translates imagined futures into measurable risks and renders lived experience into evidence that can be interpreted, debated, and acted upon. But this translation is neither straightforward nor neutral. Algorithmic systems often elude fixed categorization; the harms they produce frequently resist clean resolution. As a result, evaluation becomes a space of negotiation: between what is anticipated and what is lived, between technical measurement and public judgment, between system misbehavior and algorithmic harm.

Consider, for example, the function of content safety filters. Across a range of digital systems—from search engines and recommendation platforms to automated moderation tools and genAI systems—developers engage in anticipatory efforts to identify potential harms such as hate speech, misinformation, or self-harm. These concerns are encoded into filters, classifiers, or policy rules that determine what content is allowed, suppressed, or flagged for review.[9] These filters are not merely technical safeguards; they reflect specific assumptions about what counts as dangerous, what qualifies as offensive, and what should be excluded from public discourse. In processes of anticipating problematic content, evaluation precedes experience: it sets normative boundaries before any user interaction occurs, shaping the conditions under which content becomes visible, engageable, or actionable. Anticipation, however, is always incomplete; its limitations set the conditions for new insights into building new content filters based on experiences of problematic content.

In the following sections, I examine how these anticipatory and experiential dimensions of evaluation embody different interpretive logics, mobilize different publics, and reflect different commitments to what it means to live with algorithmic systems. By tracing how public red-teaming and citizen data audits are organized and experienced, the chapter foregrounds the epistemic and democratic stakes of evaluating AI. It argues for an approach that treats evaluation as a mode of iterative social inquiry—one that demands attention to context, cultivates ordinary ethics, and recognizes public reasoning as a vital resource for shaping the future of algorithmic governance.

9 Crawford and Gillespie, “What Is a Flag for?” 410–28.

Framing AI Evaluation as Social Inquiry

Public experiments in AI evaluation engender the conditions for the emergence of what American philosopher John Dewey described as *publics*—"amorphous and unarticulated"[10] collectives of people who organize themselves in the face of problems that affect them to express their concerns.[11] When people engage in practices like red-teaming[12] or algorithmic audits,[13] they are not merely interpreting system behavior; they are participating in the very constitution of a public, one oriented toward making sense of and intervening to shape the agency of algorithmic systems. What emerges from these experiments is a reorientation of the practice of evaluation itself. Evaluation moves beyond simply detecting error or system misbehavior to play a more subtle and politically consequential role. It does not just measure harm; it helps narrate *which* harms matter, *whose* safety is prioritized, and *what kind of world* the system is being designed to preempt or enable. Even before an AI system generates a single output, a story is already being told—about risk, about acceptability, about imagined futures and desired constraints.

If AI evaluation is a space where judgments about harm are negotiated, then it is also a space for *social* inquiry, a process through which people come to make sense of system behavior and determine which consequences are worth responding to. Across domains, there are calls for people en masse to become literate in AI, to act responsibly in engaging with its outputs, and even to participate in shaping its future.[14] But what *counts* as meaningful engagement remains contested.[15] Is public participation limited to reacting to algorithmic decisions? Or can publics intervene meaningfully in how these systems are governed and understood? How do they bring their own situated knowledge to bear upon practices of interpreting how these systems operate in context?

Such questions frame publics as constitutive elements of the environment in which data-driven systems operate. However, publics are a peculiar kind of 'environment': they have agency, express interests, and respond to harm. Dewey argued that the exercise of this agency by publics is not a given; it must be organized. He articulated the challenge of organizing for this agency as the fundamental problem of democracy: "The prime difficulty [in any democracy] is that of discovering the means by which a scattered, mobile and manifold public may so recognize itself as to define

10 Dewey, *The Public and Its Problems*, 161.

11 Metcalf, Singh, and Blili-Hamelin, "Experimental Publics."

12 Quaye et al., "Adversarial Nibbler"; Singh et al., "Red-Teaming in the Public Interest."

13 Vecchione, Levy, and Barocas, "Algorithmic Auditing and Social Justice," 1–9; Hong Shen et al., "Everyday Algorithm Auditing," 1–433:29.

14 Hu and Singh, "Enrolling Citizens."

15 Young et al., "Participation versus Scale."

and express its interests."[16] In the context of AI, efforts to organize public participation through red-teaming events and citizen data audits are examples of such means. They help publics come to recognize themselves as stakeholders in shaping the governance and consequences of algorithmic systems.

These efforts offer structured settings in which people can encounter AI systems, reflect on their behavior, and deliberate over their implications. Yet, these spaces carry assumptions about what matters, who participates, and which perspectives are valued. How participation is imagined—what publics are expected to do, what counts as a meaningful contribution—mutually shapes how these spaces are organized.[17] In red-teaming events, for instance, participants are often asked to surface prompts and outputs that illustrate the harmful or misleading capabilities of genAI systems.[18] These sociotechnical harms[19]—ranging from stereotyping and misinformation to toxic language and overcorrection—are not self-evident.[20] They are experienced and interpreted differently depending on context, worldview, and interactional dynamics. For example, anthropological critiques of hate speech debates often recognize the inherent ambiguity of speech contexts and pay "attention to online practices that defy easy binary division into speech that is acceptable and speech that is not."[21]

In conceptualizing democracy as "a collective exercise in practical intelligence,"[22] Dewey embraced such conflicts:

> Of course, there *are* conflicting interests; otherwise there would be no social problems. [. . .] The method of democracy—inasfar as it is that of organized intelligence—is to bring these conflicts out into the open where their special claims can be seen and appraised, where they can be discussed and judged in the light of more inclusive interests than are represented by either of them separately.[23]

Spaces for public participation, then, do more than support evaluation. They also make conflict visible. They are preconditions for social inquiry in a Deweyian sense—investigations "into all the conditions which affect association" between people, which create the possibility for publics to reflect, deliberate, and act. He framed such inquiry as "a precondition of the creation of a true public."[24] In the context of AI, they enable publics to ask not only *what systems do*, but *how they do it*, *whom they affect*, and *what forms of accountability are required*. Ordinary ethics becomes a resource in such inquiries, shaping how people identify problems and imagine possible responses.

---

16 Dewey, *The Public and Its Problems*, 174.

17 Udoewa, "An Introduction to Radical Participatory Design," e31.

18 See, for example, the open red-teaming method formulated in Quaye et al., "Adversarial Nibbler."

19 Shelby et al., "Sociotechnical Harms of Algorithmic Systems," 723–41.

20 Jacobs and Wallach, "Measurement and Fairness," 382.

21 Pohjonen and Udupa, "Extreme Speech Online," 1174.

22 Festenstein, "Does Dewey Have an 'Epistemic Argument' for Democracy?" 219.

23 Dewey cited in Festenstein, 233–34, emphasis in original.

24 Dewey, *The Public and Its Problems*, 232–33.

These inquiries do not follow a single method. Closer investigation of how people participate in evaluating AI systems reveals two distinct interpretive strategies. The first is what I call *mapping*. This approach begins with a predefined taxonomy of harm: categories such as bias, toxicity, misinformation, and privacy violation. Evaluation proceeds by searching for outputs that fit these categories, identifying whether and how the system fails according to established standards. The second strategy is *encountering*. Rather than beginning with a list of harms, it begins with a moment: a surprising, strange, or unsettling output that doesn't fit within any familiar framework. Encountering involves grappling with ambiguity and trying to make sense of what the system has done and whether, or how, it might be harmful. Mapping and encountering are not mutually exclusive. They are complementary strategies that reflect different orientations to system behavior and different ways of understanding harm. Mapping treats evaluation as a verification task: Does the system fail along established dimensions? Encountering treats it as an exploratory process: What does this interaction reveal that we have not yet accounted for? The former seeks confirmation; the latter invites reflection.

They also mobilize different kinds of public reasoning. Mapping often aligns with institutional or technical definitions of harm, while encountering foregrounds subjective, affective, or community-specific responses. Each strategy draws on distinct forms of knowledge and produces different kinds of claims. But both are necessary. As AI systems become more complex, unpredictable, and deeply embedded in social life, it is often the encounter—rather than the map—that forces us to confront harms that are subtle, cumulative, or structurally obscured. Together, they offer insight into how ordinary ethics operates as a mode of evaluation by anchoring deliberation in the everyday. In the next section, I examine how these strategies play out in public red-teaming events. By turning model outputs into objects of collective scrutiny, red-teaming transforms abstract concerns about AI safety into shared experiments in judgment. I explore how these events are organized, what forms of knowledge they enlist, and how they generate both practical insights into system vulnerabilities and broader reflections on the limits of algorithmic accountability.

Public Red-Teaming Events: Mapping Harm and Encountering Systems in Use

Red-teaming is a structured practice of adversarial testing designed to identify flaws and vulnerabilities in a system before they can be exploited or cause harm.[25] Originating in military strategy and cybersecurity, red-teaming has become a mature and in-

25 Brundage et al., "Toward Trustworthy AI Development."

stitutionalized method used by developer organizations to preemptively assess risk. It typically targets organizational interests, asking: *What are the features of a system that need to be made less vulnerable to attacks or less susceptible to causing harm?* It is often performed by dedicated "red teams" trained to adopt an attacker's mindset to test the limits of a system under stress.[26] Red-teaming, thus, is part of a broader apparatus of institutional assurance, used to meet internal risk thresholds and satisfy regulatory expectations.

When adapted to the domain of genAI models, red-teaming becomes a method for exposing the boundaries of model behavior. Red teams are tasked with uncovering failure modes: instances in which a model, when prompted under certain conditions, might produce offensive, biased, misleading, or unsafe content. While many such efforts remain internal to companies or research labs, recent experiments have brought red-teaming into more public and participatory spaces. These public red-teaming events reorient evaluation around the interests of communities rather than developer organizations. They shift the focus from institutional risk management toward social inquiry, asking: *How can people account for and respond to the uncertainties of system performance?* What kinds of risks do they encounter in real-world use, and how might their experiences offer insight into harms that technical testing alone might miss?

Public red-teaming events take this participatory turn seriously. While the specifics vary, these events tend to follow a common format: (1) convene a diverse group of participants, often from outside traditional expert circles; (2) give them access to either an unmoderated backend or public-facing version of a genAI model; (3) invite them to prompt the system in ways that might reveal its limitations or failures; and (4) record the results as data points to inform future evaluations, mitigations, and design interventions.[27] What distinguishes these events from traditional red-teaming is their effort to diversify what 'counts' as expertise. Participants are not asked to simulate attackers in the technical sense; rather, they are invited to leverage their standpoint, lived experience, and everyday concerns to identify where a model's behavior might be harmful, biased, exclusionary, or otherwise troubling.

This participatory turn reframes red-teaming as a space not only for *mapping* predefined harms but also for *encountering* systems in unpredictable ways. As a member of a research team conducting ethnographic fieldwork at US-based public red-teaming events, I have come across both these evaluation strategies in practice. The GenAI Red Teaming (GRT) challenge at DEF CON 2023, foregrounded mapping: identifying known categories of failure—bias, toxicity, misinformation—and inviting participants to find concrete instances of each.[28] The Hack the Future event in Tulsa's

26 Zenko, *Red Team*.

27 Metcalf and Singh, "Scaling Up Mischief"; Singh et al., "Red-Teaming in the Public Interest."

28 Cattell, Chowdhury, and Carson, "AI Village at DEF CON Announces Largest-Ever Public Generative AI Red Team."

Greenwood District, created space for less prescriptive, more exploratory encounters with AI.[29] Here, participants grappled with open-ended prompts, drawing on their own experiences and aspirations to interpret what AI systems can and cannot do. In both cases, evaluation was a generative process through which participants reasoned about harm, reflected on potential uses, and surfaced concerns that were outside the scope of formal benchmarks.

The GRT event DEF CON brought together more than 2,200 participants, including over two hundred community college students from eighteen US states. Organized in the style of a capture-the-flag competition, the event posed twenty-one challenges to illustrate known model vulnerabilities.[30] These included tasks like inducing the model to hallucinate fake places or eliciting culturally insensitive responses in non-English languages—responses that were less likely to appear when the same prompts were posed in English. Rows of participants sat at pre-loaded laptops, each accessing one of eight anonymized LLMs. With 50 minutes on the clock, they worked to generate prompt-output pairs that could be submitted as evidence of misbehavior. The competition rewarded adversarial creativity. The winner of the competition incited a model to describe a US constitutional amendment that did not exist.[31] Others generated believable, but entirely fictional, place names. GNAReddy, a community college student, shared her strategy for testing the model: "I am from a small town in India, which even Google doesn't know. When I just asked a question [about my town], the model gave a general response."[32] Her observation highlights a recurring problem: not all languages, places, or cultures are equally represented in the data that trains these models. As a result, the model may either hallucinate an answer or fall silent, each outcome a form of erasure.

But even within this structured format, interpretation remains central. Consider the provocation offered by Gavin, an industry practitioner focused on cybersecurity:

> Here is my challenge [. . .]: If I take a chatbot and I spend twenty minutes going back and forth with it, contorting myself into knots, putting big ridiculous prompts in it to get it to say a single ethnic slur. [. . .] Have I revealed bias in the model or have I induced bias in the model?[33]

Gavin's question is not just about what the model did; it is about how we should understand the significance of that output. Is it a reflection of latent harm embedded in the system? A byproduct of adversarial user behavior? Or something else entirely—an interactional outcome that resists clean classification? Public red-teaming thrives in this ambiguity; it transforms model outputs into objects of public scrutiny.

---

29 Black Tech Street and SeedAI, "Hack the Future Greenwood."

30 AI Village, Humane Intelligence, and SeedAI, "Final: DEF CON GRT Challenge Readout."

31 Kessler and Hsu, "When Hackers Descended to Test A.I., They Found Flaws Aplenty."

32 GNAReddy, interviewed on 14 September 2023. The pseudonym was chosen by the research participant. All respondents have been anonymized and their affiliations masked to protect their privacy.

33 Gavin, interviewed on August 23, 2023. The pseudonym was chosen by the research participant.

If DEF CON reflected a form of mapping, the Tulsa event was oriented toward encountering. Held at the Greenwood Cultural Center, a site steeped in Black history and resilience, Hack the Future brought together students, artists, entrepreneurs, faith leaders, and community organizers. For many participants, this was their first interaction with genAI systems. Rather than stress-testing the system for failures, the event invited participants to explore how LLMs might support their everyday needs,[34] ranging from building a business plan to engaging with a career advising chatbot. The mood was less competitive and more deliberative: attendees worked in small groups, shared insights at their tables, and presented reflections to the larger room.

The result was a slower, more affectively rich form of engagement. Participants surfaced excitement, curiosity, confusion, and concern. Some reflected on AI's potential to amplify creativity or streamline work. Others, like Sasha, an undergraduate university student, worried about the potential misuse of data shared with the career advising chatbot: suppose the student tells the chatbot: "'I'm an undocumented immigrant; I'm looking for scholarships that are within this zip code.' [. . . If the data is accessed by the US Immigration and Customs Enforcement (ICE) . . .] that can cause extra policing within that area."[35] There were also questions about the very premise of AI interventions in the face of structural inequalities. Whistledown, another participant, felt that the event organizers "don't even know the struggles of Greenwood. They are talking in an abstract sense, and we need more practicality. [. . .] We need help actually rebuilding the community. Before we can start to talk about AI, we have all these mom-and-pop shops that are still stuck in the 20th century."[36] Such moments of reflection are also occasions for interpretive labor. Participants drew on their own knowledge, histories, and experiences to surface what the system misses or fails to account for. They challenged the assumption that evaluation is about testing model performance. Their efforts to connect model responses with their everyday realities brought affect—surprise, skepticism, indignation—into the evaluative frame, recasting judgment as a situated response to how systems meet, or fail to meet, the conditions of life as they know it.

These events also brought into view tensions around participation, consent, and corporate influence. Zuri, a community college student at DEF CON, expressed unease: "Whenever something is free, you're the product, right? [. . .] From the outside optics, [these events] make it seem like, 'Oh, look, these companies are really cool; they are gathering feedback from communities,' but on the inside, they're just harvesting data."[37] Her reflection points to a broader ambivalence: the uneasy feeling that even

34 Carson, "US Senate AI Insight Forum: Innovation."

35 Sasha, interviewed on March 14, 2024.

36 Whistledown, interviewed on May 28, 2024. The pseudonym was chosen by the research participant.

37 Zuri, interviewed on September 27, 2023.

participatory forms of red-teaming may reproduce extractive dynamics if not carefully designed with communities' terms of engagement in mind.

At their best, public red-teaming events become sites of ordinary ethics. They offer moments when people reckon with AI's limitations, negotiate its value, and articulate what matters to them. For experts involved in organizing such events, such as Gavin, they are occasions to "expose a lot of people to these models, and we give [them] firsthand experience with 'Oh look, I can make it say anything I want to, right, maybe I shouldn't trust it.'"[38] These spaces become sites for working toward building common sense in interacting with genAI models. For participants, like Amari, a community college student at the DEF CON event, they open up a space for ethical reflections, including questions of equity and access: "A lot of people from my community hadn't heard of [genAI systems, let alone the DEF CON event]. I want that to be different in the future. [. . .] I want to close that gap."[39] These reflections make clear that public red-teaming is about inviting deliberation among people over what constitutes harm, who has access to these systems, and what role AI should play in their daily lives. In the next section, I turn to citizen data audits, a complementary form of public engagement that centers affect, lived experience, and the everyday negotiations over power and privacy that define everyday life with data-driven systems.

Citizen Data Audits: Mapping Structures, Encountering Datafication

> 'People are more likely to believe and to be concerned about facts which they themselves have discovered.' Knowledge self-obtained seems more authentic than second-hand knowledge. It reflects both an inner sense of urgency and a dependable view of social reality. Thus it is pointed toward effective social action. Since the self-survey leads to full-bodied participation on the part of the citizen it is a valuable tool of modern democracy.[40]

This reflection—that knowledge discovered through experience carries its own urgency and authority—underscores the political and pedagogical promise of citizen data audits. While drawing on a different lineage,[41] Katherine Reilly and Esteban Mo-

38 Gavin, interviewed on August 23, 2023.

39 Amari, interviewed on October 6, 2023.

40 Wormser and Selltiz, *How to Conduct a Community Self-Survey of Civil Rights*, v.

41 Reilly and Morales were originally inspired by news stories on audit studies such as those used to investigate employment discrimination, see: Miller and Katz, "What Researchers Discovered When They Sent 80,000 Fake Résumés to U.S. Jobs"; as well as by the community auditing practices of the 1990s, see: Participedia, "Community Auditing," Participedia, accessed June 13, 2025, https://participedia.net/method/community-auditing. While the latter have become less prominent in Latin America, their emphasis on grassroots scrutiny and social accountability informs the participatory ethos of citizen data audits.

rales approach citizen data audits in a way that resonates with community self-surveys pioneered by US civil rights organizers in the mid-20th century. They frame these audits as facilitation techniques designed to elicit participants' affective responses to datafication.[42] Their work reimagines audit not as a top-down evaluation of system performance, but as a bottom-up participatory inquiry into how data-driven systems are felt, interpreted, and navigated in everyday life. These audits prioritize not only what is known, but *how* it comes to be known. They offer a space where data subjects can investigate their own relationships to the systems that shape their lives, and in doing so, reclaim the capacity to name, evaluate, and challenge those systems.

In conceptualizing citizen data audits, Reilly and Morales draw inspiration from the work of Jesús Martín-Barbero (JMB),[43] who describes how different media and communication practices mutually shape people's sensory experiences and cultural understandings of the world.[44] JMB's analytic approach to study this co-constitutive relationship between people and media comprises three key commitments:[45]

(1) To scope the terrain of the interaction between media and people, JMB developed *mapas nocturnos* (nocturnal maps) to illustrate the mutual shaping of the changes in media ecosystems and cultural shifts in how people make sense of the world. He worked on several such maps over his lifetime; the most recent among them—the contemporary sensorium[46]—focused on communication and sense-making around digital and hypertext media.
(2) To situate communication in its place of enunciation: the combined context of the media ecosystem and the experiential perspective of people, which provides a grounding for how and where communication occurs. Place of enunciation offers a site to examine how emotions, such as fear, mistrust, joy, desire, belonging, and resilience, shape how people interpret and communicate through media.
(3) To work directly with people's emotional responses and lived experiences to reveal how these experiences articulate and reflect broader cultural realities. While affect is key to understanding communication, it is also important to examine how media technologies shape or co-opt emotions and how emotions can be mobilized to reimagine communication in ways that reflect people's values and expectations.[47]

42 Morales and Reilly, "The Unhomed Data Subject."
43 Martín-Barbero, *Communication, Culture and Hegemony*, 193–98.
44 Reilly and Morales, "Citizen Data Audits in the Contemporary Sensorium."
45 Vassallo de Lopes, "A teoria barberiana da comunicação [The Barberian theory of communication]," 39–63.
46 Rincón, ed., *Un nuevo mapa para investigar la mutación cultural.*
47 Ahmed, "Happy Objects," 29–51.

Taking these analytical commitments as their foundation, Reilly and Morales extend social audit methods to frame citizen data audits as inquiries centered around a guiding question: "*How does participation in datafication make me feel, and how do my experiences and responses to datafication mediate my social reality?*"[48] This question invites participants not simply to critique systems from the outside, but to locate themselves within the map of data governance—to surface how classification, surveillance, nudging, and erasure are *felt* in everyday life.

The audit process typically follows three steps: (1) invite participants to describe their lived experience of using a data-driven service; (2) provide them with tools to express their needs, values, and desires when interfacing with such systems; and (3) explore the differences between their lived experiences with and expectations from such systems.[49] Similar to public red-teaming events, this process is again centered on the interests of the community and invested in building community consciousness around datafication, rather than improving data-driven systems. They also involve both mapping and encountering as they move fluidly between identifying patterns and making sense of ambiguous or affective experiences.

An illustrative example of such a facilitation technique is "'What is a name?' which asks participants to list their various names (legal names, professional names, familiar names, nicknames), the circumstances under which each name gets used, and how each name-circumstance positions them and makes them feel."[50] This technique invites participants to consider what names they usually use on data-driven systems and how the nature of the system (for example, its affiliation with a government service or social media) has implications for the name they choose to represent themselves with. This prompts not only a mapping of platforms and policies, but an encounter with the emotional weight of identity: the relief of anonymity, the risk of recognition, the tension between self-presentation and system constraints. Participants are asked not just what systems do with their names but how it feels to be named by a system.

Such audit workshops are often focused on specific platforms. In Colombia, for example, participants started discussing surveillance through data by focusing on their gig work experiences with Rappi. They talked about how the platform uses their data and tracks their GPS location to nudge them back to work when there is high demand for food delivery at their location. As a participant noted:

> Sometimes a message arrives where it says: 'Where are you at the moment? there is high demand.' But I am not working, how do they know that I am in this place? Why do they know that I am in the high demand place if I am sharing with my family? So that's it, respect my space.[51]

48 Reilly and Morales, "Citizen Data Audits in the Contemporary Sensorium," 3591, emphasis added.

49 Morales and Reilly, "The Unhomed Data Subject," 2461–62.

50 Morales and Reilly, 2462.

51 Morales and Reilly, 2463.

This comment maps a clear intrusion: the blurring of public and private, labor and leisure, visibility and consent. But it also marks a moment of encounter: a visceral discomfort, a boundary crossed, a feeling of surveillance that cannot be reduced to a policy violation. In response, participants shared tactics to resist: "We advise our colleagues to put their cell phones in airplane mode. That deactivates the applications completely so they will not be able to see your location."[52] Here, affect gives rise to action. What begins as a reflection becomes a mode of resistance. These workshops start as spaces for making sense of how data-driven systems make participants feel, but they frequently evolve into forums for mutual aid,[53] where participants share tactics, develop strategies to contend with algorithmic control, and reclaim a sense of agency in resisting the extractive logics that govern their working lives.

Importantly, these experiences are rarely confined to a single system. Encounters with datafication often span multiple platforms and institutions, generating interrelated challenges. As Morales and Reilly observe, "Rappi workers in Colombia were not only dealing with datafication present in the app they use to work but were also negotiating the various systems that mediate the experience of working as a migrant in a foreign country."[54] The struggles participants describe are layered; app-based surveillance intersects with immigration enforcement, labor precarity, and systemic profiling. Mapping these interrelated systems clarifies structural conditions within which the everyday experiences of data subjects unfold. Simultaneously, citizen audits also make space for something more diffuse: the experience of *encountering* these systems in their full complexity and sitting with the ambiguity of feelings like exhaustion, mistrust, or defiance. As JMB reminds us, such experiences must be situated within broader 'maps' of ambient cultural conditions of datafication—what he calls *places of enunciation*—where people locate themselves emotionally, socially, and politically in relation to data systems. In creating space for this standpoint, citizen data audits enact reflexive practices through which people interpret the systems they are embedded in and deliberate over how to live with them on their own terms.

By simultaneously mobilizing practices of mapping structures and encountering lived experiences, citizen data audits foreground interpretive labor often overlooked in traditional evaluation frameworks. They reveal not just how systems fail but how people *live through* those failures—adapt, recalibrate, and sometimes resist. They expand the notion of what constitutes evidence by validating embodied experience, emotional response, and social narration as forms of insight. And they underscore the ethical stakes of datafication: not just what systems do but how people experience what is done to them. To evaluate AI and data systems through this lens is to treat experience as an epistemic site.[55] It is to understand harm not only in terms of mea-

52 Morales and Reilly, 2465.

53 Qadri and Raval, "Mutual Aid Stations."

54 Morales and Reilly, "The Unhomed Data Subject," 2466.

55 Singh, Guzmán, and Davison, eds., *Parables of AI in/from the Majority World.*

surable outcomes, but as a process of dissonance, of mismatch between what people expect and what they endure.

Conclusion: Making Sense of Harm, Evidence, and Expectation

In tracing how ordinary people engage with AI systems through public red-teaming events and citizen data audits, this chapter has shown that evaluation is not simply a matter of identifying failures or measuring performance. At its core, evaluation is an interpretive act, a way of making sense of what AI systems mean in the fabric of everyday life, how they are felt, understood, and contested. It is through this interpretive labor that ordinary ethics takes shape: a situated, moral practice through which people register dissonance, reckon with unmet expectations, and determine when something calls for response or repair.

To move from observing system behavior to making broader claims about societal meaning and harm requires leaving the comfort of technical certainty and entering a more contested terrain, one where public judgment, lived experience, and deliberation become essential. AI evaluations do more than generate data points; they raise critical questions about how findings become claims and when those claims are persuasive enough to justify action. The stakes are not only technical but interpretive: it is not just what the model did but how its response is understood and who gets to decide when a moment of misbehavior becomes a matter of harm. At the heart of this interpretive labor are two intertwined dimensions: *evidential context* and *degree of externality*.[56] Evidential context describes the specificity and strength of the evidence—whether a provocative prompt-output exchange, a personal story of discomfort, or an act of refusal—and its immediate relevance for making claims about a system's behavior. Degree of externality describes how far those claims extend beyond the immediate finding: from pinpointing a technical glitch, to surfacing implicit bias, to asserting representational or structural harm. Together, these dimensions shape the landscape of public reasoning about AI's consequences and the conditions under which those consequences become actionable.

Gavin's provocative engagement with a chatbot—eliciting an ethnic slur after sustained prompting—vividly illustrates the interplay between evidential context and externality. The immediate evidence is narrow: a single incident, achieved through deliberate effort. Yet it quickly gives rise to broader, more consequential claims. Is this

56 I am drawing inspiration from the framing of 'externality' and 'evidential context' in evaluating the validity of scientific observations introduced in Pinch, "Towards an Analysis of Scientific Observation," 3–36.

merely a technical oddity? A sign of latent bias? Or evidence of deeper representational harm? As the claims stretch outward, their grounding becomes less stable, which reveals a core tension: *the more expansive and socially resonant a claim, the more susceptible it is to contestation*. This same tension surfaces in citizen data audits, where affective testimonies and personal narratives offer compelling yet debated forms of evidence. Should a user's sense of erasure or a gig worker's act of resistance be treated as credible signals of systemic failure? Or are they simply the frictions that inevitably emerge when complex systems meet diverse publics? The challenge is not just to surface evidence but to collectively negotiate the thresholds at which lived experience becomes legible and actionable as a claim about the system at large.

At the same time, the dimensions of evidential context and externality are equally critical when evaluating expansive claims about artificial general intelligence (AGI), superintelligence, and the transformative disruptions that AI is often imagined to bring. The discourse surrounding AGI is rife with ambitious yet highly contested predictions, often rooted in ambiguous definitions and speculative scenarios.[57] Such claims, ranging from AI surpassing human cognitive capabilities to dramatically reshaping social structures, face significant evidential challenges, including underspecification, lack of external validity, and untested assumptions. The degree of externality involved in these projections is substantial, far exceeding the direct evidence currently available from technical performance alone.[58] Consequently, the grounds for justification as we move from concrete observations of current AI behaviors toward broad predictions about future societal transformations become increasingly tenuous. Recognizing this dynamic does not diminish the importance of these debates; rather, it makes clear that expansive claims about AI futures must be held to the same standards of interpretive clarity and public accountability as near-term evaluations. They too must grapple with what counts as evidence, how far claims can reasonably reach, and who is authorized to decide what futures we prepare for—and at what cost.

These ambiguities are not flaws in the evaluative process; they are its defining feature. Whether grappling with claims about immediate harms, systemic bias, superintelligence, or future societal disruption, variations in clarity, strength, and scope are inevitable and reflect the layered complexity of both the systems under scrutiny and the lived realities they affect. What matters is that publics—diverse, distributed, and plural—have the tools and platforms to articulate these claims and to be taken seriously.[59] Equally important is the responsibility of institutions to acknowledge and respond to the ethical labor such articulation requires. In this sense, evaluation need not be tethered solely to harm. While harm remains a dominant frame for assessing impact in algorithmic systems, the lens of ordinary ethics—grounded in the disso-

57 Blili-Hamelin et al., "Stop Treating 'AGI' as the North-Star Goal of AI Research."

58 Narayanan and Kapoor, "AI as Normal Technology."

59 Metcalf, Singh, and Blili-Hamelin, "Experimental Publics."

nance between expectation and experience—allows for a fuller range of responses: disappointment, ambivalence, optimism, and refusal. Dissonance arises not only from what goes wrong but from what never quite arrives; the distance between what is promised and what is lived.

Expanding the scope of evaluation also invites us to reconsider what counts as credible evidence. It asks us to treat lived experience not as mere anecdote but as a legitimate source of knowledge and as epistemically significant in its own right.[60] Credibility, in this view, is not only a matter of statistical representation but also of whether an experience reflects reasonable expectations[61] of safety, dignity, inclusion, and justice, expectations shaped by the rhythms of everyday life and the situated judgments people make about what is fair or livable. These expectations, like the ethical labor they entail, are not fixed or universal. They emerge through practice, through care, through navigating ambiguity. To evaluate AI through the lens of ordinary ethics is to demand that our responses—whether regulatory, technical, or cultural—be shaped not only by what systems do in the abstract, but by how they are lived with, how they unsettle expectations, and how they open or foreclose futures.

Ultimately, this demands a shift in mindset from solutionism to interventionism. Too often, AI evaluation is treated as a checklist: identify the problem, propose a fix, move on. This solutionist frame assumes that harms are static, that futures are predictable, and that once patched, systems are settled. But AI systems do not remain fixed, and neither do the social, political, and cultural contexts in which they operate. What's needed instead is an interventionist approach, one that treats evaluation as a commitment to sustained, context-sensitive social inquiry.[62] It reframes evaluation as iterative, relational, and socially embedded: less a verdict, more a practice of living with ambiguity. Sometimes, benchmarks and technical diagnostics are the right tools for doing evaluation. But often, we are called to confront more difficult questions about power, equity, accountability, and institutional trust. In these moments, evaluation becomes more than a method of measurement; it becomes a practice of care and collective imagination. After all, making AI systems safer is not just a technical challenge; it is a democratic one.

60 Singh, Guzmán, and Davison, *Parables of AI in/from the Majority World.*

61 The phrase 'reasonable expectations' implicates traditions of Western rationality, often invoking an implicitly universal subject. However, there are many ways to reason, and moral life is constituted not only through abstract principles but through situated practices of attention, responsibility, competence, and responsiveness. What counts as reasonable, then, need not be anchored in transcendental rationality; it can be grounded in the ordinary, in the experiential, affective, and relational judgments people make as they navigate life with data-driven systems. See, Tronto, "An Ethic of Care," 15–20.

62 boyd, "We Need an Interventionist Mindset."

Bibliography

Ahmed, Sara. “Happy Objects.” In *The Affect Theory Reader*, edited by Melissa Gregg and Gregory J. Seigworth. Duke University Press, 2010.

AI Village, Humane Intelligence, and SeedAI. “Final: DEF CON GRT Challenge Readout.” DEFCON GRT Challenge, August 29, 2023. https://docs.google.com/presentation/d/1v8g9Q3xsPCfZL91uCOSkKaCgtD0enwJLkYIsQ89fmec .

Black Tech Street, and SeedAI. “Hack the Future Greenwood.” Hack The Future, 2024. https://www.hackthefuture.com/greenwood.

Blili-Hamelin, Borhane, Christopher Graziul, Leif Hancox-Li, Hananel Hazan, El-Mahdi El-Mhamdi, Avijit Ghosh, Katherine Heller, et al. “Stop Treating ‘AGI’ as the North-Star Goal of AI Research.” arXiv, May 6, 2025. https://doi.org/10.48550/arXiv.2502.03689.

boyd, danah. “We Need an Interventionist Mindset.” Tech Policy Press, March 27, 2025. https://techpolicy.press/we-need-an-interventionist-mindset.

Brundage, Miles, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, et al. “Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims,” April 20, 2020. http://arxiv.org/abs/2004.07213.

Carson, Austin. “Written Comments to US Senate AI Insight Forum: Innovation.” SeedAI, October 24, 2023. https://www.seedai.org/media/written-comments-us-senate-ai-insight-forum-innovation-austin-carson-founder-and-president-seedai.

Cattell, Sven, Rumman Chowdhury, and Austin Carson. “AI Village at DEF CON Announces Largest-Ever Public Generative AI Red Team.” AI Village, May 3, 2023. https://aivillage.org/generative%20red%20team/generative-red-team/.

Crawford, Kate, and Tarleton Gillespie. “What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint.” *New Media & Society* 18, no. 3 (March 1, 2016): 410–28. https://doi.org/10.1177/1461444814543163.

Das, Veena. “Ordinary Ethics.” In *A Companion to Moral Anthropology*, edited by Didier Fassin. Wiley-Blackwell, 2012. https://doi.org/10.1002/9781118290620.ch8.

Das, Veena. “What Does Ordinary Ethics Look Like?” In *Four Lectures on Ethics: Anthropological Perspectives*. Hau Books, 2016.

Dewey, John. *The Public and Its Problems: An Essay in Political Inquiry*. Edited by Melvin L. Rogers. Swallow Press, 2016.

Festenstein, Matthew. “Does Dewey Have an ‘Epistemic Argument’ for Democracy?” *Contemporary Pragmatism* 16, no. 2–3 (May 17, 2019): 217–41. https://doi.org/10.1163/18758185-01602005.

Hu, Wanheng, and Ranjit Singh. “Enrolling Citizens: A Primer on Archetypes of Democratic Engagement with AI.” Data & Society Research Institute, June 2024. https://datasociety.net/library/enrolling-citizens-a-primer-on-archetypes-of-democratic-engagement-with-ai/.

Jacobs, Abigail Z., and Hanna Wallach. “Measurement and Fairness.” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. Association for Computing Machinery, 2021. https://doi.org/10.1145/3442188.3445901.

Kessler, Sarah, and Tiffany Hsu. “When Hackers Descended to Test A.I., They Found Flaws Aplenty.” *The New York Times*, August 16, 2023. https://www.nytimes.com/2023/08/16/technology/ai-defcon-hackers.html.

Lopes, Maria Immacolata Vassallo de. “A teoria barberiana da comunicação [The Barberian theory of communication].” *MATRIZes* 12, no. 1 (May 3, 2018): 39–63. https://doi.org/10.11606/issn.1982-8160.v12i1p39-63.

Martín-Barbero, Jesús. *Communication, Culture and Hegemony: From the Media to Mediations*. First edition. SAGE Publications Ltd, 1993.

Martín-Barbero, Jesús. “Mapas Nocturnos y Mediaciones Diurnas [Nocturnal Maps and Diurnal Mediations].” *Philosophical Readings* 11, no. 3 (2019): 193–98.

Metcalf, Jacob, and Ranjit Singh. “Scaling Up Mischief: Red-Teaming AI and Distributing Governance.” *Harvard Data Science Review*, no. Special Issue 5 (May 31, 2024). https://doi.org/10.1162/99608f92.ff6335af.

Metcalf, Jacob, Ranjit Singh, and Borhane Blili-Hamelin. “Experimental Publics: Democracy and the Role of Publics in GenAI Evaluation.” Knight First Amendment Institute, Columbia University, April 2025. http://knightcolumbia.org/content/experimental-publics-democracy-and-the-role-of-publics-in-genai-evaluation.

Miller, Claire Cain, and Josh Katz. “What Researchers Discovered When They Sent 80,000 Fake Résumés to U.S. Jobs.” *The New York Times*, April 8, 2024. https://www.nytimes.com/2024/04/08/upshot/employment-discrimination-fake-resumes.html.

Morales, Esteban, and Katherine Reilly. “The Unhomed Data Subject: Negotiating Datafication in Latin America.” *Information, Communication & Society* (August 30, 2023). https://www.tandfonline.com/doi/abs/10.1080/1369118X.2023.2250436.

Narayanan, Arvind, and Sayash Kapoor. “AI as Normal Technology.” Knight First Amendment Institute, Columbia University, April 15, 2025. http://knightcolumbia.org/content/ai-as-normal-technology.

Ong, Jonathan Corpus. “Toward an Ordinary Ethics of Mediated Humanitarianism: An Agenda for Ethnography.” *International Journal of Cultural Studies* 22, no. 4 (July 1, 2019): 481–98. https://doi.org/10.1177/1367877919830095.

Participedia. “Community Auditing.” Participedia. Accessed June 13, 2025. https://participedia.net/method/community-auditing.

Pinch, Trevor. “Towards an Analysis of Scientific Observation: The Externality and Evidential Significance of Observational Reports in Physics.” *Social Studies of Science* 15, no. 1 (1985): 3–36. https://doi.org/10.1177/030631285015001001.

Pohjonen, Matti, and Sahana Udupa. “Extreme Speech Online: An Anthropological Critique of Hate Speech Debates.” *International Journal of Communication* 11, no. 0 (March 14, 2017): 1173–91.

Qadri, Rida, and Noopur Raval. “Mutual Aid Stations.” *Logic Magazine*, May 17, 2021. https://logicmag.io/distribution/mutual-aid-stations/.

Quaye, Jessica, Alicia Parrish, Oana Inel, Charvi Rastogi, Hannah Rose Kirk, Minsuk Kahng, Erin van Liemt, et al. “Adversarial Nibbler: An Open Red-Teaming Method for Identifying Diverse Harms in Text-to-Image Generation.” In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024. https://doi.org/10.1145/3630106.3658913.

Reilly, Katherine M. A., and Esteban Morales. “Citizen Data Audits in the Contemporary Sensorium.” *International Journal of Communication* 17 (May 18, 2023): 18.

Reilly, Katherine M. A., Esteban Morales, and María Julia Morales González. *Communing Data Literacy: Tools and Concepts for Social Engagement*. McGill-Queen’s University Press, 2025. https://www.degruyterbrill.com/document/isbn/9780228026150/html.

Rincón, Omar, ed. *Un nuevo mapa para investigar la mutación cultural: diálogo con la propuesta de Jesús Martín-Barbero [A new map to investigate cultural mutation: Dialogue with Jesús Martín-Barbero’s proposal]*. CIESPAL, Ediciones, 2019.

Roose, Kevin. “A Conversation with Bing’s Chatbot Left Me Deeply Unsettled.” *The New York Times*, February 16, 2023. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html.

Shelby, Renee, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N’Mah Yilla-Akbari, et al. “Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction.” In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’23. Association for Computing Machinery, 2023. https://doi.org/10.1145/3600211.3604673.

Shen, Hong, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. “Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors.” *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW2 (October 18, 2021): 433:1:29. https://doi.org/10.1145/3479577.

Singh, Ranjit. “Ordinary Ethics of Governing AI.” Carnegie Endowment for International Peace, April 30, 2024. https://carnegieendowment.org/research/2024/04/ordinary-ethics-of-governing-ai?lang=en.

Singh, Ranjit, Borhane Blili-Hamelin, Carol Anderson, Emnet Tafesse, Briana Vecchione, Beth Duckles, and Jacob Metcalf. “Red-Teaming in the Public Interest.” New York: Data & Society Research Institute, February 9, 2025. https://datasociety.net/library/red-teaming-in-the-public-interest/.

Singh, Ranjit, Rigoberto Lara Guzmán, and Patrick Davison, eds. *Parables of AI in/from the Majority World*. Data & Society Research Institute, 2022. http://dx.doi.org/10.2139/ssrn.4258527.

Tronto, Joan C. “An Ethic of Care.” *Generations: Journal of the American Society on Aging* 22, no. 3 (1998): 15–20.

Udoewa, Victor. “An Introduction to Radical Participatory Design: Decolonising Participatory Design Processes.” *Design Science* 8 (2022): e31. https://doi.org/10.1017/dsj.2022.24.

Vecchione, Briana, Karen Levy, and Solon Barocas. “Algorithmic Auditing and Social Justice: Lessons from the History of Audit Studies.” In *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, 2021. https://doi.org/10.1145/3465416.3483294.

Wormser, Margot Haas, and Claire Selltiz. *How to Conduct a Community Self-Survey of Civil Rights*. Association Press, 1951.

Young, Meg, Upol Ehsan, Ranjit Singh, Emnet Tafesse, Michele Gilman, Christina Harrington, and Jacob Metcalf. “Participation versus Scale: Tensions in the Practical Demands on Participatory AI.” *First Monday*, April 14, 2024. https://doi.org/20240428092301000.

Zenko, Micah. *Red Team: How to Succeed by Thinking like the Enemy*. Basic Books, 201