Sébastien A. Krier using Midjourney 6.1

ESSAYS AND SCHOLARSHIP

Experimental Publics: Democracy and the Role of Publics in GenAI Evaluation

Public involvement both challenges and reshapes the governance of generative AI systems.

BY JACOB METCALF, RANJIT SINGH & BORHANE BLILI-HAMELIN
APRIL 8, 2025

ARTIFICIAL INTELLIGENCE AND DEMOCRATIC FREEDOMS

A project studying how advanced AI systems may harm, or help strengthen, democratic freedoms

Abstract

This paper examines how emerging public experiments in red-teaming generative AI (genAI) systems engender new formations of 'experimental publics,' shaped by competing visions of expertise, democratic oversight, and AI safety. Interrogating the practice of genAI red-teaming, the authors analyze distinctions between instrumental and deliberative public feedback, situating them within the historical Dewey-Lippmann debate on democratic governance. By bridging AI evaluation with political epistemology, they explore how public involvement both challenges and reshapes the governance of genAI systems, revealing its democratic potential as well as its contested boundaries.

Introduction

> **Walter Lippmann**: The interest of the public is not in the rules and contracts and customs themselves but in the maintenance of a regime of rule, contract and custom. The public is interested in law, not in the laws; in the method of law, not in the substance; in the sanctity of contract, not in a particular contract; in understanding based on custom, not in this custom or that. [...] The pressure which the public is able to apply through praise and blame, through votes, strikes, boycotts or support can yield results only if it reinforces the men who enforce an old rule or sponsor a new one that is needed.[1]

> **John Dewey**: No government by experts in which the masses do not have the chance to inform the experts as to their needs can be anything but an oligarchy managed in the interests of the few. And the enlightenment must proceed in ways which force the administrative specialists to take account of the needs. [... This] depends essentially upon freeing and perfecting the processes of inquiry and of dissemination of their conclusions. Inquiry, indeed, is a work which devolves upon experts. But their expertness is not shown in framing and executing policies, but in discovering and making known the facts upon which the former depend. [...] It is not necessary that the many should have the knowledge and skill to carry on the needed investigations; what is required is that they

> have the ability to judge of the bearing of the knowledge supplied by others upon common concerns.[2]

Arguments over the role of the public in democratic governance have animated debates among political philosophers for nearly a century. These arguments—as represented in the distinct positions taken by Walter Lippmann and John Dewey in the quotes above—are centered on answering two core questions: (1) what should be the proper relationship between citizens and experts in a world overflowing with information yet anchored in democratic values, and (2) under what conditions can the public come together to realize self-governance? Both these questions are about the constitutive dimensions of governing with distributed expertise. On the one hand, Lippmann believed that it is impossible to democratically govern in a complex modern information environment where the ordinary public is expected to be informed and adequately exhibit expertise in a wide variety of topics.[3] As a prominent journalist and World War I propagandist, he was deeply concerned about the rise of radio and mass market print journalism creating an information environment that exceeded the capacity of a normal citizen to consume and responsibly interpret without being manipulated, and therefore, he favored a technocratic form of government managed by experts. On the other hand, Dewey, the foremost American pragmatist philosopher of his day, believed that the complexity of information environments is the condition of possibility for "amorphous and unarticulated" collectives of people to organize themselves in the face of problems that affect them to express their concerns.[4] He saw democracy as an ongoing experiment, one where policies are continuously tested and refined through deliberations over their lived experience, and where expertise is measured not by the ability to frame policies, but by the capacity to reveal the underlying facts upon which they depend.

In this paper, we examine elements of the Lippmann-Dewey debate to analyze the tensions in democratic governance of generative AI (genAI) and its broader implications for AI safety. The fundamental imbalance of power between the average citizen and the tech companies that can afford to spend billions of dollars to build foundation models would appear to be a nearly-Platonic ideal of a 'technocracy,' where a class of technical experts is empowered to impose its peculiar vision of the future of human life upon a politically-impotent polis, fully setting the terms of a general-purpose decision engine without robust public input.[5] Yet many of these same developers are cautiously embracing public feedback as a necessity for the good functioning of their technical systems, a distinctly Dewey-

esque position. Embedded within these different positions are distinct expectations around the role that the public writ large can play in organizing and governing AI safety. We map these expectations by focusing on the practice of genAI red-teaming as the empirical site where debates over approaches that involve the public and expectations of the role that public participation can play in evaluating genAI are being actively debated.

In 2023, genAI reached an unprecedented level of public visibility, driven largely by the release of ChatGPT by OpenAI in late 2022. These systems, built on flexible models trained on vast datasets, generate new media in response to plain-language prompts. As their capabilities became more apparent, a diverse array of actors—including academics, security professionals, and the public writ large—began actively probing their weaknesses. By identifying problematic 'prompt-output pairs,' they demonstrated how genAI systems could produce harmful or unintended results. This emerging form of public interrogation paralleled the rise of internal AI governance and safety teams within tech companies and research labs, often described as 'AI red-teaming.' Borrowing a term from cybersecurity, red-teaming refers to systematic efforts to expose a system's vulnerabilities. Across multiple jurisdictions, regulatory agencies have increasingly positioned red-teaming as a fundamental, if not obligatory, component of the safe development and deployment of genAI systems.[6]

What distinguishes genAI red-teaming from traditional security red-teaming efforts, however, is the growing recognition that what counts as acceptable behavior of a genAI system should not be solely decided by employees of tech companies—an insight acknowledged by those same employees.[7] As a result, many genAI red-teaming experiments seek public participation in different forms, from engineering prompts to assessing harmful outputs. We examine genAI red-teaming as a critical intersection between 'the public' and 'experts.' Having traditionally relied on cybersecurity expertise, this technical practice is evolving to incorporate a wide range of lay public knowledge and lived experience as essential components of genAI governance.

Borrowing Dewey's framing of experimentalist democracy,[8] we frame *experimental publics* as a conceptual resource to articulate this emergent type of public participation in technology governance broadly and genAI governance specifically. We define *experimental publics* as emergent and provisional collectives assembled through structured interventions in making sense of sociotechnical systems—particularly at moments when system uncertainties and decision stakes are high,

and the traditional distinction between facts and values becomes difficult to maintain.[9] At such junctures, the boundaries of public participation, technical authority, and governance are unsettled and actively contested. While related to frameworks like co-design, citizen science, and participatory AI, experimental publics are distinct in that they take shape around organized activities of public evaluation—epistemic exercises that both generate knowledge and reorganize relations of authority. Unlike participatory approaches that assume stable contributory roles for participants in design or governance, experimental publics foreground instability, provisionality, and the generative uncertainty of participatory evaluation itself.[10] We analyze genAI red-teaming as a site for new formations of experimental publics—redefining how systems are evaluated, how authority is distributed, and how public accountability is enacted.

Such public red-teaming experiments received significant attention in May 2023 when the White House announced participation of leading large language model (LLM) developers in a public Generative Red Team (GRT) event at DEF CON, the world's largest annual computer security conference.[11] We conducted participant observation at this event and studied other ongoing efforts throughout 2023 and 2024 to organize public participation in red-teaming genAI systems. These efforts include, but are not limited to:

1. The Adversarial Nibbler challenge (launched in July 2023), which crowdsourced diverse failure modes to evaluate the safety of text-to-image genAI models. [12]

2. The AI Democracy Project (held in January 2024), which piloted an expert-driven, domain-specific safety testing of five leading genAI language models, focusing on their responses to election misinformation.[13]

3. A purple-teaming event (organized in February 2024) in Greenwood, Tulsa, Oklahoma, inspired by the DEF CON gathering but emphasizing collaboration over adversarial testing. This event combined "red-teaming exercises with experiential real-world use case exploration in key areas of Black life."[14]

These experiments highlight the evolving relationship between professional expertise and public engagement in computer security that undergirds broader public participation in evaluating genAI. Not only do they extend the historical trajectory of efforts to incorporate public input into security work, but they also serve as a means of reckoning with the ethical dimensions and emergent consequences of genAI's proliferation in everyday life. Furthermore, they often

bring together a diverse range of stakeholders, including—but not limited to—model developers, government officials, representatives from non-profit organizations, academics, and members of the broader public.

We argue that this diversity in event design and underlying priorities reflects a deeper multiplicity in conceptions of publics and their interests. Informed by the struggles our interview participants expressed over the nature of their expertise, we contend that debates surrounding AI governance broadly, and red-teaming in particular, reveal a crucial insight: *there is no singular 'public interest' that can be easily defined or measured against*. Instead, what emerges is an assemblage of many publics, each with distinct concerns, perspectives, and stakes—disclosed and shaped through the formation of *experimental publics*. Interrogating these public experiments as forms of red-teaming, we ask: Who constitutes these publics? Why should they be involved in red-teaming? What is the nature of their contribution to red-teaming genAI systems? How should the process of soliciting their contribution be organized? We draw on a corpus of interviews with professional, amateur and educational AI red-teaming practitioners (n=28) to explore how different publics come to be involved in genAI governance processes, and how their different forms of expertise inform what we do and do not know about making these systems safe.

Practitioners we interviewed usually imagined two distinct roles for such public feedback in AI safety. Both these roles are essential to serving the public interest, but they differ in emphasis. The first, which we term instrumental public feedback, views feedback as a tool for improving systems.[15] Here, our use of the term 'instrumental' draws its lineage from the crucial role that instruments play in scientific experiments in measuring, observing, and analyzing data, testing hypotheses, and confirming theories.[16] In this sense, public feedback serves as an instrument to argue for and validate improvements to systems. The second, which we call deliberative public feedback, considers feedback as an intrinsic component of public oversight. Without public feedback, the oversight process would not be public. This is the kind of feedback associated with the place of public opinion, freedom of speech, civil society, and institutions like the free press in democratic oversight. We do not consider one form of public feedback as inherently superior to the other; rather, each plays a distinct and complementary role in the broader project of democratic AI governance.

To further develop this distinction, we draw on Bruno Latour's work in *Give Me a Laboratory and I Will Raise the World* where he argued that the power of the laboratory lies not just in its capacity to isolate, but to translate—to render the

vastness of the world into micro-scale abstractions, stable enough to study and act upon. Yet this translation is incomplete until those abstractions are scaled back up, until they move and matter in the world again. The ethnographer's task, Latour suggests, is to follow how scientists manage this movement of scale.[17] We extend this insight to the terrain of public feedback. Instrumental feedback scales down public concerns—translating diffuse anxieties into simplified abstractions that can be used to improve systems such as metrics (e.g. policy violation rate, false refusal rate), datasets, or reports. Deliberative feedback, by contrast, scales up from the technical to the social—transforming an abstract prompt-output pair into a public discussion about priorities, experiences, and ways to hold those in power to account. Each mode addresses a distinct problem of scale, and together they are complementary tools that form the scaffolding for a more participatory approach to technical governance. In what follows, we examine how genAI red-teaming experiments mobilize these modes of feedback, and how they are reshaping debates about democratic governance, public expertise, and the ongoing challenge of building safe and accountable AI systems.

Ongoing Debates Over the Role of the Public and Its Role in GenAI Evaluation

In August 2023, we began our research project on genAI red-teaming with participant observation at the public genAI red team (GRT) event at DEF CON.[18] Our goal was to investigate the intersection of red-teaming, sociotechnical safety, and public participation. As our exploration unfolded, we discovered a diverse array of experiments in organizing public participation in red-teaming ranging from competitions and bounty programs to educational events and focus group discussions.[19] The more we examined the varied designs of these events and the priorities shaping them, the more we found echoes of long-standing debates around the role of the public in AI governance. Specifically, these contemporary discussions resonate with the broader debate on 'political epistemology' at the messy intersection of information, democracy, technology, and expertise—an inquiry that animated American public philosophy nearly a century ago, most notably through the Lippmann-Dewey debate.[20]

On one hand, Walter Lippmann, a prominent social critic and journalist, argued in his books *Public Opinion* and *The Phantom Public* that robust democratic engagement on major issues of collective governance was largely hopeless due to the complexity of modern life. Due to the information environment created by radio, mass print media, and new forms of travel and communication, the citizens

of democracies were "saddled with an impossible task and … asked to practice an unattainable ideal [of direct democratic participation]. … I cannot find time to do what is expected of me in the theory of democracy; that is, to know what is going on and to have an opinion worth expressing on every question which confronts a self-governing community."[21] Lippmann's skepticism about the plausibility of collective democratic governance stemmed from his perspective (informed in part by his early career as a World War I propagandist) that a coherent collection of citizens called 'the public' was illusory, and democracy was too reliant upon the whims of people prone to manipulation and incapable of digesting the firehose of mass media information. This perspective reflected widespread sentiments among America's intellectual and political elite about the credibility of liberal democracy following World War I. Instead, Lippmann favored a quasi-democratic technocracy governed by 'scientific management' that could rise above the unreliable and irrational subjectivity that struggled to cope with industrialization, bureaucracy, and modern communications.

On the other hand, John Dewey rejected the idea that deliberating publics should be imagined as a "universal discussion of the people," or as divorced from the "identities, interests, and needs" that differentiate communities. [22] He conceptualized publics as manifestations of "amorphous and unarticulated" collectives of people who organize themselves in the face of problems that affect them to express their concerns. [23] Dewey argued that the exercise of this agency by public(s) is not a given; it must be self-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 and express its interests."[24] Furthermore, he emphasized that conflicting interests among plural publics are far from inimical to democracy. Rather, these temporary publics focused on shared concerns are essential to democracy's ability to collectively identify legitimate solutions to social problems:

> 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.[25]

Spaces to organize public participation are a crucial condition for bringing conflicts out into the open. They are a key feature of Dewey's conception of social inquiry "into all the conditions which affect association" between people and are organized through public discussions grounded in exchange of perspectives; he framed such social inquiry as "a precondition of the creation of a true public."[26] The process of "giving citizens control over the forces that govern and enable their lives" requires active collaboration between publics, experts, and institutions in social inquiries that shape and organize collective answers to contested social problems.[27]

The contemporary landscape of genAI governance offers a vivid stage for the reemergence of Lippmann and Dewey's concerns. Lippmann's skepticism about public expertise resonates with AI governance regimes that privilege expert audits, closed evaluation pipelines, and regulatory sandboxing—where accountability flows upward to regulators but not outward to publics. By contrast, Dewey's emphasis on publics as emergent and situated aligns with efforts to invite publics not only to critique AI systems, but in shaping how harm, failure modes, and ethical use are defined. Public red-teaming experiments operate squarely within this tension: Is 'the public' merely a collective of non-experts and end-users, excluded from the design of highly complex systems like genAI models and left to navigate their potential harms? Or does the public bring valuable perspective and experiential knowledge of living with such systems—knowledge that no number of computer science degrees can reconstitute, yet is a necessary component of any successful system? In the following sections, we explore these questions through the lens of securing *instrumental* as well as *deliberative* public feedback. Framing public red-teaming through the lens of political epistemology foregrounds the question of how publics come to recognize themselves as stakeholders in the safety and governance of genAI. As Emily, an industry practitioner focused on AI safety, put it when reflecting on her experience of co-organizing a public red-teaming competition: "Safety is inherently something that needs to be defined continuously by a broad range of people. So it made sense to put that in the form of a public competition."[28] These experiments, then, are not simply mechanisms for technical evaluation—they are invitations to collective inquiry, where communities are enlisted to define what counts as problematic behavior, and to imagine how it might be addressed.

Instrumental Public Feedback

Instrumental public feedback has become increasingly central to the evaluation of computational systems, particularly as traditional security frameworks—primarily

focused on technical exploits—have struggled to address emergent harms such as disinformation, trolling, harassment, and extremist content. These issues, amplified by platform design and algorithmic curation, turned content moderation into a crucial public concern throughout the 2010s.[29] The vast scale of user-generated content on social media platforms created conditions in which the public, as both end-users and impacted communities, became integral to security considerations—not only as targets of harm but also as potential contributors to platform safety. Security and product safety efforts increasingly relied on user-generated reports of vulnerabilities and flagged content, integrating them into the process of identifying potential abuses. Companies, in turn, began developing automated techniques informed by this data to detect and filter harmful content. Yet these interventions have proven insufficient. Content moderation remains an ongoing challenge, with companies outsourcing much of the labor to crowd workers in the majority world to reduce costs associated with the relentless, round-the-clock monitoring of content.[30] The evolving nature of harms associated with user-generated content highlights that the vulnerabilities of technologies that leverage user participation—particularly social media platforms—are not merely technical but sociotechnical in nature.[31] As a result, the public's role in these platforms has expanded beyond merely generating content; users have been implicitly enrolled as enforcers of security and moderators of content, shaping the very mechanisms through which platforms identify and address misuse.

The current push to engage the public in genAI red-teaming builds on this trajectory but introduces a *crucial shift*. With the public release of genAI models, traditional security efforts—once centered on protecting computer systems from exploits, managing user-generated content, and developing automated techniques to detect harmful material—must now also address the risks posed by content generated by the models themselves. As Sam, an industry practitioner focused on responsible AI, explained to us:

> The real thing we're grappling with is the transition from using AI primarily for discriminative tasks, where you're trying to label content or rank things versus using it to generate text and imagery. [...] It is no longer just user-generated content; it is [company] generated content. [There is a need for a] higher standard [...] for content that is generated by a [company] model, which can be perceived as carrying [the company's] voice.[32]

At the heart of public interventions in genAI red-teaming is the challenge of grappling with the unpredictability of model behavior—the core problem that these efforts are designed to address.

Major industry actors developing genAI models have largely treated instrumental public feedback as a means of refining or generating micro-scale and abstract rulesets that shape the content and behavior of already-built systems. For example, researchers at DeepMind have proposed the SocioTEchnical Language agent Alignment (STELA) method, which employs small, intensive focus groups drawn from underrepresented communities for community-based rule elicitation.[33] Inspired in part by the Design from the Margins methodology, STELA seeks to develop a general ruleset by inviting a small number of people from "differently situated" groups historically excluded from social power.[34] They found that starting from the perspective of marginalized groups elicited a distinct set of fairness principles. In contrast, researchers at Anthropic have pursued a participatory governance model called "constitutional AI," which structures model constraints much like a constitution sets limits on government authority.[35] This approach solicits broad crowdsourced input to again refine a proposed AI constitution based on global human rights documents, which then informs the reinforcement learning stage of model training.[36] Meanwhile, OpenAI has experimented with grant-making initiatives to explore "democratic inputs to AI," funding various projects that translate public or lay-expert input into rulesets in the form of "value cards" that developers can deploy.[37] Despite their differences, these approaches share a common goal: *scaling down and encoding public feedback into rules to govern powerful, centralized AI models*. In each case, instrumental public feedback serves two primary functions: (1) affording their governance practices a degree of external input from the broader public, and (2) refining technical safety measures as code.

Instrumental public feedback is increasingly seen as a vital mechanism for assessing both the harms and benefits of genAI systems across diverse communities. The extraordinary flexibility of these models—often described as their general-purpose nature—expands the range of contexts in which practitioners must grapple with uncertainties surrounding their societal impact. As Arvind Narayanan observes, "traditionally in ML, building models is the central activity and evaluation is a bit of an afterthought. But the story of ML over the last decade is that models are more general-purpose and more capable. General purpose means you build once but have to evaluate everywhere."[38] This "build once, evaluate everywhere" mindset is crucial for making instrumental public feedback particularly valuable for evaluating genAI models. By incorporating perspectives

from different communities, public red-teaming efforts are often designed to leverage such feedback to help surface potential risks and contextual challenges that might otherwise go unnoticed. Initiatives such as Adversarial Nibbler, the DEF CON Generative Red Team Event, and the AI Democracy Project all integrate instrumental public feedback as a key component of their design, positioning it as an essential tool in the evolving landscape of AI evaluation.

However, these efforts also introduce a set of emerging concerns. Kabir, an industry practitioner focused on machine learning ethics and policy, pointed out that most public-facing genAI portals incorporate three standard reporting mechanisms—thumbs up, thumbs down, and a report button—which mirror content-flagging systems on social media. While he acknowledged that this "is probably a decent way of collecting information," he also noted a critical limitation: "We don't know how that translates into model retraining. [...] There is a lot of intransparency surrounding these things."[39] Beyond issues of transparency, independent red teamers who operate "in the wild," outside of sanctioned red-teaming events, face potential legal risks when disclosing vulnerabilities.[40] In the US, there are no legal protections for researchers conducting good-faith testing on genAI models. As a result, they risk account suspension, legal action, or even lawsuits for violating terms of service.[41] Legal risks aside, the immediate public disclosure of vulnerabilities can inadvertently create or exacerbate security and safety risks, both for model developers and the broader public. One potential solution comes from security research: coordinated vulnerability disclosure (CVD), championed by CERT/CC, offers a structured approach to handling high-risk genAI flaws.[42] By notifying system owners in advance and providing them time to address issues before public disclosure, CVD can potentially help balance the need for transparency with responsible risk mitigation.

Deliberative Public Feedback

Less frequently discussed in the genAI evaluation space is the role of deliberative public feedback—feedback cultivated through societal conversations about the role of technology broadly and genAI particularly in everyday life. This form of engagement serves multiple purposes, including building public consciousness around genAI systems, interrogating power structures, and reflecting on the ways in which the public can meaningfully contribute to genAI development. *The core question in such discussions is whether power is exercised for the right ends, in the right way, and by the right people?*[43] When it comes to interrogating power, the human, financial, and environmental costs of genAI systems have heightened long-

standing concerns about concentration of power in technology companies.[44] Silicon Valley has long wielded disproportionate influence over AI tools that shape surveillance, search, criminal justice, hiring, health, and ad targeting. Along similar lines, the UN AI Advisory Body has warned that a major obstacle to public-interest AI governance is the imbalance of power between a handful of countries shaping AI policy and the vast majority left with little influence.[45] Foundation models further exacerbate these concerns, as a small set of companies now produce models that could be deployed across countless industries, use cases, and geographies.

The framing of genAI governance around 'safety' and 'risk management' carries both practical utility and conceptual limitations. While safety and risk management paradigms provide a language for managing direct harms such as toxic content or model misbehavior, they must be publicly interrogated for how they obscure more contested and ambiguous questions of control and responsibility, including structural inequality, systemic displacement, and epistemic harm.[46] From a Dewey-inspired perspective, similar considerations apply to the work of governments. While often organized around solving social problems, this work must be supplemented—and at times reoriented—by public deliberation. As with past technologies, such as industrial machinery that displaced workers while improving production efficiency, some harms emerge not as failures of design but as features of broader political-economic arrangements. Deliberative public feedback offers a crucial corrective by surfacing concerns that may fall outside the scope of traditional safety metrics, urging attention not only to model behavior, but to the *social consequences* of its adoption. A compelling instance of such deliberative engagement emerges from artistic communities that have mobilized protest movements against genAI-generated art. Through tactics such as public "call-outs," "banning or flagging AI-generated art on art-sharing sites," "data poisoning," and "legal action," these communities have pushed back against the commercialization of AI art.[47] A shared concern among these protests is the argument that AI-generated art constitutes a form of theft—that developers have illegitimately appropriated artists' works to train models for commercial use, without consent or compensation.

Creating conditions for such deliberation, however, is no simple task. Samantha, an expert in AI risk management and safety standards who also helped organize a community red-teaming event, articulated a common underlying expectation in these initiatives: "the public does not need to be involved in the nitty-gritty details of how AI is designed and developed. They don't care; they just want to be

protected. To the extent that red teamers can provide that protection, that is incredibly valuable."[48] Public engagement in AI governance is often framed through a sense of safety—people feel protected as they gain opportunities to experiment with different forms of participation. These approaches include building familiarity with genAI systems (AI literacy), exercising agency in response to model misbehavior (AI governance), and contributing to model development (Participatory AI).[49] Each offers a distinct foundation for public deliberation, shaping how communities navigate and contest the evolving role of genAI in their everyday lives.

Furthermore, these approaches raise questions around the very premise of organizing public red-teaming events. Traditionally, red-teaming has been designed around the interests of organizations that are building technical systems, focusing on improving the security and safety of those systems; the societal benefits of robust security red-teaming are secondary effects. It typically asks: *What vulnerabilities need to be addressed to make a system more resilient to attacks or less likely to cause harm?* However, novel experiments in public red-teaming shift the focus toward community interests, reframing the core question: *How can the public assess and respond to a system's uncertainties, enabling more cautious and informed use?* This shift is evident in qualitative studies of red-teaming in the wild, where researchers identified a key distinction: professional red teamers "are explicitly looking for 'failure modes'" of genAI models, while individuals engaged in jailbreaking "are often looking to get the model to obey."[50] This distinction highlights an adversarial exercise of agency—users pushing models to misbehave not just as an intellectual challenge, but as a way to test and exploit their boundaries. Beyond their ability to manipulate these systems, publics are increasingly recognized as cultural experts and key stakeholders in defining what AI harms look like. They are positioned as both the first to experience and the last line of defense against the consequences of model misbehavior. When end-user populations engage responsibly with genAI models, they contribute not only to lowering the likelihood of AI harm but also to shaping the broader discourse on AI safety and governance.

This prioritization of people, rather than systems, is central to how deliberative public feedback is organized within community red-teaming efforts, which typically: (1) emphasize competitions that reward expertise in identifying diverse failure modes, and (2) raise awareness and provide educational opportunities to help participants understand how genAI systems can fail. Beyond their potential to contribute to AI safety and the specter of embarrassing public incidents such as the

controversy over images showing multi-racial Nazi-era German soldiers generated by Google's Gemini in February 2024, both organizers and participants see community red-teaming as more than just a technical exercise.[51] These events are framed as opportunities to engage with a like-minded community and deepen public understanding of genAI systems to ultimately achieve a different shared purpose: *scaling up public conversations on potential consequences of genAI*. As one participant of the GRT event at DEF CON put it, "I get to be around all these extremely intelligent people and learn so much."[52] At the same time, they also function as competitive spaces where individuals can demonstrate their expertise in uncovering system flaws—as Samantha put it, "There is always going to be some person who is going to perform, if you have these challenges, way better than anybody else."[53] This dual nature of community red-teaming—both as a collaborative learning space and a competitive testing ground—left some participants with mixed feelings. For instance, Zuri, a community college student, reflected on the tension she felt about her participation in such a competition:

> I have two perspectives on it. One, I think there are a couple of reasons why it's good. Because you have the justice aspect of changing power dynamics, and [ensuring] equitable development of AI. We have diverse data sets, open evaluation. I also see it from a negative standpoint, because I was worried about data privacy. [... By participating] you're giving something away that is valuable to [developers], whether you know it or not. People may not have the knowledge or consent even available to participate in some of these studies and would unwillingly or unknowingly give away their information, because that's way more valuable than their feedback. Because from the outside optics, [community red-teaming events] makes it seem like, "Oh, look, these companies are really cool." They're gathering feedback from communities, but on the inside, they're just harvesting data, which is way more [...] useful to them.[54]

Despite these mixed feelings, raising awareness about genAI models remained central to the organization of community red-teaming events, particularly in discussions around the ethics of access. Amari, reflecting on the DEF CON event, highlighted this concern:

> A lot of people from my community hadn't heard of [the DEF CON event] or didn't know anything about it. I want that to be different in the future. I know how important it is because [of the difference between] the kids that are on a chatbot now at [age] eight versus the Black community, who might be anywhere else. It [will] have large effects; in a year's time that eight year old could be doing this and that, get all of his homework done, and have all of this extra free time. I want to close that gap [for my community].[55]

Beyond education as an outcome, community red-teaming events serve as spaces to explore and contest ethical concerns that matter to the public. Students engaged not only with issues of fairness and bias but also demonstrated an intuitive grasp of how and why bias manifests in genAI models. As Amari continued: "Yes, the Black community is definitely harmed by [...] AI because we have the least amount of data. AI is all based on data. So if you have [less] data about a person, you will find that [these systems] easily make mistakes [...] because [they] do not have enough data to know that that's offensive or incorrect."[56] These events also function as sites for crystallizing critical thinking about model behavior, providing participants with the tools to interrogate AI systems and their broader societal impacts.

Finally, the expectation that communities will engage with genAI models can sometimes clash with their readiness to do so, particularly when it comes to identifying and understanding model misbehavior. Returning to Samantha's earlier point, the public does not need to be involved in every technical detail of AI design and development. The key question is *when*, *who*, and *how* to involve members of the public in these processes while balancing expectations around their contributions. Some participants felt that red-teaming is a learned skill and that community red-teaming events should not only measure participants' interest in genAI but also assess their level of understanding to ensure meaningful engagement. As one participant of the purple teaming event at Greenwood, Tulsa, reflected:

> It is not a one size fits all [event. Participants should be] incentivized to actually explore more. But in a healthier sense, it's not like information overload. [...] For me just being in that room [during the purple-teaming event in Greenwood, the skill level of participants] was like easily level zero to level one. [... When you're only at a beginner level, such an event

> can be overwhelming.] It's kind of hard to give caviar to a baby. I feel like that's what was happening there. Give the baby milk, then give them soft foods and fruit, then move to solid, then change the portion size. So I just think that there's a way to do it. We can't rush it. [...] You can't impact anything, if you don't even know what your impacts are and the thing to impact are people, like people are really your focus group. If we can't understand that, what are we doing?[57]

Slowly building expertise among the wider public and creating space for communities to explore and articulate their interests and concerns is central to organizing deliberative public feedback through community red-teaming. These events reflect distinct expectations of public participation, including: (1) becoming familiar with and learning how to use genAI in everyday life; (2) engaging in collective decision-making processes to address the emerging consequences of genAI; and (3) drawing on personal experiences, expertise, and perspectives to contribute to AI design and to identify failure modes in genAI systems.[58] These expectations, in turn, shape the methods used to organize public participation and the nature of the deliberative feedback emerging from these events. For developers, the value of community red-teaming lies in generating instrumental public feedback—from red-teaming datasets to a broader understanding of subjective targets of evaluation, such as diverse conceptions of AI harms and deeper insights into the contexts of normal, instead of adversarial, use of genAI models.[59] For members of the public, these events create spaces for deliberation—allowing them to interact with genAI systems, learn from one another, develop their own understanding of the systems' workings and limits, and, at times, engage directly with AI and domain experts.

To conclude our analysis of public feedback, the case studies we followed reveal how experimental publics operate across a spectrum of instrumental and deliberative feedback. The *Adversarial Nibbler Challenge* exemplifies instrumental feedback: it solicits targeted failure modes to refine system behavior, emphasizing adversarial rigor and detailed annotations of harm. By contrast, the *Greenwood purple-teaming event* centers deliberative feedback, positioning red-teaming as a mode of community reflection, ethical inquiry, and collective engagement with genAI systems. The DEF CON GRT event occupies a hybrid space, blending structured expert-led evaluation categories with open public participation—staging a dialogue between technical benchmarks and diverse civic concerns. Mapping these events against our conceptual framework highlights how the design of

participatory infrastructure—its format, affordances, and framing—shapes the epistemic and political character of public involvement.[60]

Conclusion

While red-teaming has become a prominent method for evaluating genAI systems, it is only one among many overlapping mechanisms of AI governance. Evaluation—especially when it includes public participation—intersects with audit regimes, transparency requirements, impact assessments, and broader regulatory frameworks. The role of experimental publics, then, is not merely to test AI models, but to expand the scope of what counts as valid input into how systems are shaped, judged, and legitimized. Public evaluation can complement expert audits by surfacing situated harms, elevating contestation, and challenging assumptions embedded in technical design. It also exposes the limits of existing regulatory logics, particularly when participatory feedback exceeds the narrow boundaries of what current AI risk management regimes define as 'safety.' Integrating evaluation into a more pluralistic governance ecosystem demands attention to these tensions—and to the epistemic and political stakes they carry.

Along these lines, Seth Lazar, drawing on democratic ideals of freedom, equality, and collective self-determination, argues that the exercise of power in AI must satisfy a strong *publicity* requirement: "reasonably competent members of the governed community must be able to determine that they are being governed legitimately and with proper authority."[61] In this framework, publicity imposes obligations on those in power, including the duty to explain their decision-making to their political community. Similarly, Reuben Binns suggests that the democratic ideal of public reason could help navigate reasonable disagreements over governance norms. He cautions that accountability efforts—which enable meaningful scrutiny of whether AI governance actors fulfill their legal, political, or societal obligations—risk being undermined by reasonable disagreement over what constitutes legitimate governance.[62] Drawing on the political philosophy tradition, Binns advocates for grounding algorithmic accountability in "common principles" that citizens can broadly accept, rather than in "controversial propositions which citizens might reasonably reject."[63]

This challenge of legitimacy in decision-making within democratic governance lies at the heart of the Lippmann-Dewey debate over the roles of the public and experts in organizing for democracies. Their positions are not in binary opposition but rather represent differing views on the primacy of publics and experts along a spectrum of how decision-making processes are organized—from self-governance,

where decisions are made collectively, to technocracy, where decision-making is entrusted to experts. Navigating this spectrum requires both *scaling down* public concerns into micro-scale representations that experts can reconcile and *scaling up* opportunities for deliberation, allowing the public to engage with the macro-scale consequences of these decisions. In functioning democracies, this interplay between scaling down and scaling up is fundamental to sustaining public experimentalism.

Yet this push and pull between scaling down and scaling up public participation in decision-making processes becomes even more fraught and contested in reckoning with the uncertainties inherent in complex systems.[64] As Sydney Dekker warns, under the pressures of shifting environments, unpredictable technology, and social normalization of growing risks, stakeholders may become increasingly detached from governance processes—allowing systems to drift into failure.[65] Dewey cautions that democratic institutions face a similar threat when we act,

> as if our democracy were something that perpetuated itself automatically; as if our ancestors had succeeded in setting up a machine that solved the problem of perpetual motion in politics.[66]

Just as "validation of minority opinion and an encouragement of dissent" are vital for countering safety failures in technical systems, Dewey reminds us that processes that enable "deliberation," "discursive contestation," and "critical feedback" are essential safeguards against democratic failures.[67] This insight is particularly urgent as AI governance risks veering toward technocratic or even authoritarian control. *While scientific expertise is foundational to modern democracy, democracy itself is foundational to scientific advancement*. Together they engender the conditions of possibility for public experimentalism. Those who recognize the necessity of robust public input in shaping genAI systems—including the computer scientists designing and deploying them—must actively seek methods and infrastructures to foster experimental publics that generate both instrumental and deliberative feedback.

Bibliography

Ahmed, Shazeda, Klaudia Jaźwińska, Archana Ahlawat, Amy Winecoff, and Mona Wang. "Field-Building and the Epistemic Culture of AI Safety." *First Monday* 29, no. 4 (April 1, 2024). https://doi.org/10.5210/fm.v29i4.13626.

Anderson, Elizabeth. “The Epistemology of Democracy.” *Episteme* 3, no. 1–2 (June 2006): 8–22. https://doi.org/10.3366/epi.2006.3.1-2.8.

Anthropic. “Challenges in Red Teaming AI Systems,” June 12, 2024. https://www.anthropic.com/news/challenges-in-red-teaming-ai-systems.

———. “Claude’s Constitution.” *Anthropic Blog* (blog), May 9, 2023. https://www.anthropic.com/news/claudes-constitution.

———. “Collective Constitutional AI: Aligning a Language Model with Public Input.” *Anthropic Blog* (blog), October 17, 2023. https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input.

Arvind Narayanan @random_walker. “Traditionally in ML, Building Models Is the Central Activity and Evaluation Is a Bit of an Afterthought. But the Story of ML over the Last Decade Is That…” *X.Com*, September 8, 2024. https://x.com/random_walker/status/1840731490239340896.

Bai, Yuntao, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, et al. “Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback.” arXiv, April 12, 2022. http://arxiv.org/abs/2204.05862.

Bergman, Stevie, Nahema Marchal, John Mellor, Shakir Mohamed, Iason Gabriel, and William Isaac. “STELA: A Community-Centred Approach to Norm Elicitation for AI Alignment.” *Scientific Reports* 14, no. 1 (March 19, 2024): 6616. https://doi.org/10.1038/s41598-024-56648-4.

Binns, Reuben. “Algorithmic Accountability and Public Reason.” *Philosophy & Technology* 31, no. 4 (December 2018): 543–56. https://doi.org/10.1007/s13347-017-0263-5.

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

Burrell, Jenna, and Jacob Metcalf. “Introduction for the Special Issue of ‘Ideologies of AI and the Consolidation of Power’: Naming Power.” *First Monday* 29, no. 4 (April 1, 2024). https://doi.org/10.5210/fm.v29i4.13643.

Bybee, Carl. “Can Democracy Survive in the Post-Factual Age?: A Return to the Lippmann-Dewey Debate about the Politics of News.” *Journalism & Communication Monographs* 1, no. 1 (March 1, 1999): 28–66. https://doi.org/10.1177/152263799900100103.

Carson, Austin. “Written Comments | U.S. 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/.

Cattell, Sven, Avijit Ghosh, and Lucie-Aimée Kaffee. “Coordinated Flaw Disclosure for AI: Beyond Security Vulnerabilities.” In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7:267–80, 2024. https://doi.org/10.1609/aies.v7i1.31635.

Collective Intelligence Project. “The Collective Intelligence Project Whitepaper.” Collective Intelligence Project, 2023. https://cip.org/whitepaper.

Collins, H. M. *Gravity’s Kiss: The Detection of Gravitational Waves*. 1st edition. Cambridge: MIT Press, 2017.

Dekker, Sidney. *Drift into Failure: From Hunting Broken Components to Understanding Complex Systems*. Farnham ; Burlington, VT: Ashgate Pub, 2011.

Delgado, Fernando, Stephen Yang, Michael Madaio, and Qian Yang. “The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice.” In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–23. EAAMO ’23. New York, NY, USA: Association for Computing Machinery, 2023. https://doi.org/10.1145/3617694.3623261.

Desai, Deven R, and Joshua A Kroll. “Trust but Verify: A Guide to Algorithms and the Law.” *Harv. JL & Tech.* 31 (2017): 1.

Dewey, John. “(1939) Creative Democracy—The Task before Us.” In *The Pragmatism Reader: From Peirce through the Present*, edited by Robert B. Talisse and Scott F. Aikin. Princeton, NJ Oxford: Princeton University Press, 2011.

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

Eloundou, Tyna, and Teddy Lee. "Democratic Inputs to AI Grant Program: Lessons Learned and Implementation Plans." *OpenAI* (blog), January 16, 2024. https://openai.com/blog/democratic-inputs-to-ai-grant-program-update.

European Parliament. "The Act Texts." EU Artificial Intelligence Act, April 16, 2024. https://artificialintelligenceact.eu/the-act/.

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.

Fraser, Nancy. "Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy." *Social Text*, no. 25/26 (1990): 56. https://doi.org/10.2307/466240.

Funtowicz, Silvio O., and Jerome R. Ravetz. "Science for the Post-Normal Age." *Futures* 25, no. 7 (September 1, 1993): 739–55. https://doi.org/10.1016/0016-3287(93)90022-L.

Goerzen, Matt, Elizabeth Anne Watkins, and Gabrielle Lim. "Entanglements and Exploits: Sociotechnical Security as an Analytic Framework," 2019. https://www.usenix.org/conference/foci19/presentation/goerzen.

Goetze, Trystan S. "AI Art Is Theft: Labour, Extraction, and Exploitation: Or, On the Dangers of Stochastic Pollocks." In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 186–96. FAccT '24. New York, NY, USA: Association for Computing Machinery, 2024. https://doi.org/10.1145/3630106.3658898.

Grant, Nico. "Google Chatbot's A.I. Images Put People of Color in Nazi-Era Uniforms." *The New York Times*, February 22, 2024, sec. Technology. https://www.nytimes.com/2024/02/22/technology/google-gemini-german-uniforms.html.

Harrington, Christina N. "The Forgotten Margins: What Is Community-Based Participatory Health Design Telling Us?" *Interactions* 27, no. 3 (April 17, 2020): 24–29. https://doi.org/10.1145/3386381.

Householder, Allen D, Garret Wassermann, Art Manion, and Chris King. “The CERT Guide to Coordinated Vulnerability Disclosure.” Special Report. CMU/SEI-2017-SR-022 CERT Division, August 2017. https://resources.sei.cmu.edu/asset_files/specialreport/2017_003_001_503340.pdf.

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

Huang, Saffron, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. “Collective Constitutional AI: Aligning a Language Model with Public Input.” In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1395–1417. FAccT ’24. New York, NY, USA: Association for Computing Machinery, 2024. https://doi.org/10.1145/3630106.3658979.

Humane Intelligence. “Algorithmic Bias Bounty Programs.” Humane Intelligence, 2024. https://www.humane-intelligence.org/bias-bounty.

Inie, Nanna, Jonathan Stray, and Leon Derczynski. “Summon a Demon and Bind It: A Grounded Theory of LLM Red Teaming in the Wild.” arXiv, November 13, 2023. https://doi.org/10.48550/arXiv.2311.06237.

Latour, Bruno. “Give Me a Laboratory and I Will Raise the World.” In *Science Observed: Perspectives on the Social Study of Science*, edited by Karin Knorr-Cetina and Michael Mulkay, 141–70. London and Beverly Hills: Sage, 1983. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.473.5243&rep=rep1&type=pdf.

Lazar, Seth. “Legitimacy, Authority, and Democratic Duties of Explanation.” arXiv, October 11, 2023. http://arxiv.org/abs/2208.08628.

———. “Power and AI: Nature and Justification.” In *The Oxford Handbook of AI Governance*, edited by Justin Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew Young, and Baobao Zhang. Oxford University Press, 2022. https://doi.org/10.1093/oxfordhb/9780197579329.013.12.

Leveson, Nancy. *An Introduction to System Safety Engineering*. Cambridge, MA: The MIT Press, 2023.

Lippmann, Walter. *The Phantom Public*. 1st edition. New Brunswick, NJ: Routledge, 1993.

Longpre, Shayne, Sayash Kapoor, Kevin Klyman, Ashwin Ramaswami, Rishi Bommasani, Borhane Blili-Hamelin, Yangsibo Huang, et al. "Position: A Safe Harbor for AI Evaluation and Red Teaming." In *Proceedings of the 41st International Conference on Machine Learning*, edited by Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, 235:32691–710. Proceedings of Machine Learning Research. PMLR, 2024. https://proceedings.mlr.press/v235/longpre24a.html.

Mansbridge, Jane. "Feminism and Democracy - The American Prospect." *The American Prospect*, February 19, 1990. https://prospect.org/civil-rights/feminism-democracy/.

Melissa Heikkilä. "This New Data Poisoning Tool Lets Artists Fight Back against Generative AI." *MIT Technology Review*, October 23, 2023. https://www.technologyreview.com/2023/10/23/1082189/data-poisoning-artists-fight-generative-ai/.

Mody, Cyrus C. M. *Instrumental Community: Probe Microscopy and the Path to Nanotechnology*. Cambridge: MIT Press, 2011.

Ojewale, Victor, Ryan Steed, Briana Vecchione, Abeba Birhane, and Inioluwa Deborah Raji. "Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling." arXiv, March 14, 2024. http://arxiv.org/abs/2402.17861.

OpenAI. "Democratic Inputs to AI." *OpenAI* (blog), May 25, 2023. https://openai.com/index/democratic-inputs-to-ai/.

Ovadya, Aviv. "'Generative CI' through Collective Response Systems." arXiv, February 1, 2023. https://doi.org/10.48550/arXiv.2302.00672.

Proof News, and The Science, Technology, and Social Values Lab at the Institute for Advanced Study. "The AI Democracy Projects." Proof News, June 25, 2024. https://www.proofnews.org/tag/the-ai-democracy-projects/.

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." arXiv, May 13, 2024. https://doi.org/10.48550/arXiv.2403.12075.

Rigot, Afsaneh. “Design from the Margins.” *Harvard Belfer Center for Science and International Affairs*, 2022. https://www.belfercenter.org/sites/default/files/files/publication/TAPP-Afsaneh_Design%20From%20the%20Margins_Final_220514.pdf.

Roberts, Sarah T. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Illustrated edition. New Haven: Yale University Press, 2019.

Rogers, Melvin L. “Revisiting The Public and Its Problems.” In *The Public and Its Problems: An Essay in Political Inquiry*. Athens, Ohio: Swallow Press, 2016.

Rumman Chowdhury. “What the Global AI Governance Conversation Misses.” *Foreign Policy*, September 19, 2024. https://foreignpolicy.com/2024/09/19/ai-governance-safety-global-majority-internet-access-regulation/.

Seger, Elizabeth, Aviv Ovadya, Divya Siddarth, Ben Garfinkel, and Allan Dafoe. “Democratising AI: Multiple Meanings, Goals, and Methods.” In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 715–22. AIES ’23. New York, NY, USA: Association for Computing Machinery, 2023. https://doi.org/10.1145/3600211.3604693.

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/.

Spring, Jonathan M., April Galyardt, Allen D. Householder, and Nathan VanHoudnos. “On Managing Vulnerabilities in AI/ML Systems.” In *New Security Paradigms Workshop 2020*, 111–26. Online USA: ACM, 2020. https://doi.org/10.1145/3442167.3442177.

Subramonian, Arjun, Vagrant Gautam, Dietrich Klakow, and Zeerak Talat. “Understanding ‘Democratization’ in NLP and ML Research.” arXiv, June 17, 2024. http://arxiv.org/abs/2406.11598.

United Nations and AI Advisory Body. “Governing AI for Humanity.” United Nations, September 2024. https://www.un.org/en/ai-advisory-body.

Warner, Michael. *Publics and Counterpublics*. New York: Zone Books, 2005.

White House. “FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation That Protects Americans’ Rights and Safety.” The White House, May 4, 2023. https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/.

———. “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.” The White House, October 30, 2023. https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

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.

Cite as: Jacob Metcalf, Ranjit Singh, and Borhane Blili-Hamelin, Experimental Publics: Democracy and the Role of Publics in GenAI Evaluation, 25-09 Knight First Amend. Inst. (Apr. 8, 2025), **https://knightcolumbia.org/content/experimental-publics-democracy-and-the-role-of-publics-in-genai-evaluation**[**https://perma.cc/TH2Q-JU8Y**].

---

1 Walter Lippmann, *The Phantom Public*, 1st edition (New Brunswick, NJ: Routledge, 1993), 95.

2 John Dewey, *The Public and Its Problems: An Essay in Political Inquiry*, ed. Melvin L. Rogers (Athens, Ohio: Swallow Press, 2016), 225.

3 Lippmann, *The Phantom Public*.

4 Dewey, *The Public and Its Problems: An Essay in Political Inquiry*, 161.

5 Burrell, Jenna, and Jacob Metcalf. “Introduction for the Special Issue of ‘Ideologies of AI and the Consolidation of Power’: Naming Power.” *First Monday* 29, no. 4 (April 1, 2024). https://doi.org/10.5210/fm.v29i4.13643; Ahmed, Shazeda, Klaudia Jaźwińska, Archana Ahlawat, Amy Winecoff, and Mona Wang. “Field-Building and the Epistemic Culture of AI Safety.” *First Monday* 29, no. 4 (April 1, 2024). https://doi.org/10.5210/fm.v29i4.13626.

6 The White House, “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,” The White House, October 30, 2023,

https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence; European Parliament, "The Act Texts," EU Artificial Intelligence Act, April 16, 2024, https://artificialintelligenceact.eu/the-act/.

7 Ranjit Singh et al., "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/.

8 Dewey, *The Public and Its Problems: An Essay in Political Inquiry*.

9 We thank Henry Farrell for observing that our emphasis on the provisional character of experimental publics diverges from more familiar accounts of publics in theories of democracy, which are often understood to have some amount of stability or persistence over time. Many public red-teaming efforts have the provisional character of gatherings. However, as we note in discussing deliberative feedback, we consider the extent to which persistent community interests become centered within or emerge from such efforts remains an open question. See, Silvio O. Funtowicz and Jerome R. Ravetz, "Science for the Post-Normal Age," *Futures* 25, no. 7 (September 1, 1993): 739–55, https://doi.org/10.1016/0016-3287(93)90022-L for a detailed account of such situations where the distinction between facts and values becomes difficult to maintain in technoscientific practice.

10 Fernando Delgado et al., "The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice," in *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23 (New York, NY, USA: Association for Computing Machinery, 2023), 1–23, https://doi.org/10.1145/3617694.3623261; Meg Young et al., "Participation versus Scale: Tensions in the Practical Demands on Participatory AI," *First Monday*, April 14, 2024, https://doi.org/20240428092301000; Wanheng Hu and Ranjit Singh, "Enrolling Citizens: A Primer on Archetypes of Democratic Engagement with AI" (New York: Data & Society Research Institute, June 2024), https://datasociety.net/library/enrolling-citizens-a-primer-on-archetypes-of-democratic-engagement-with-ai/.

11 White House, "FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation That Protects Americans' Rights and Safety," The White House, May 4, 2023, https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/.

12 Jessica Quaye et al., "Adversarial Nibbler: An Open Red-Teaming Method for Identifying Diverse Harms in Text-to-Image Generation" (arXiv, May 13, 2024), https://doi.org/10.48550/arXiv.2403.12075.

13 Proof News and The Science, Technology, and Social Values Lab at the Institute for Advanced Study, "The AI Democracy Projects," Proof News, June 25, 2024, https://www.proofnews.org/tag/the-ai-democracy-projects/.

14 Austin Carson, "Written Comments | U.S. 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.

15 Saffron Huang et al., "Collective Constitutional AI: Aligning a Language Model with Public Input," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24 (New York, NY, USA: Association for Computing Machinery, 2024), 1395–1417, https://doi.org/10.1145/3630106.3658979; OpenAI, "Democratic Inputs to AI," *OpenAI* (blog), May 25, 2023, https://openai.com/index/democratic-inputs-to-ai/; Arjun Subramonian et al., "Understanding 'Democratization' in NLP and ML Research" (arXiv, June 17, 2024), http://arxiv.org/abs/2406.11598.

16 H. M. Collins, *Gravity's Kiss: The Detection of Gravitational Waves*, 1st edition (Cambridge: MIT Press, 2017); Cyrus C. M. Mody, *Instrumental Community: Probe Microscopy and the Path to Nanotechnology* (Cambridge: The MIT Press, 2011).

17 Bruno Latour, "Give Me a Laboratory and I Will Raise the World," in *Science Observed: Perspectives on the Social Study of Science*, ed. Karin Knorr-Cetina and Michael Mulkay (London and Beverly Hills: Sage, 1983), 141–70, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.473.5243&rep=rep1&type=pdf.

18 The "red-teaming in the public interest" project is a collaboration between Data & Society Research Institute (D&S) and AI Risk and Vulnerability Alliance (ARVA). Ranjit Singh led the D&S team working on the project consisting of Emnet Tafesse, Briana Vecchione, and Jacob Metcalf. Borhane Blili-Hamelin led the ARVA team that also included Carol Anderson, and Beth Duckles. The project's findings are published separately in the form of a report. See, Singh et al., "Red-Teaming in the Public Interest." This paper is based on ongoing conversations between Jake Metcalf, Ranjit Singh, and Borhane Blili-Hamelin over the role of the public writ large in genAI evaluation; all views expressed in this paper are their own and do not represent the views of the rest of the research team. Sven Cattell, 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/.

19 Quaye et al., "Adversarial Nibbler"; Humane Intelligence, "Algorithmic Bias Bounty Programs," Humane Intelligence, 2024, https://www.humane-intelligence.org/bias-bounty; Black Tech Street and SeedAI, "Hack the Future Greenwood," Hack The Future, 2024, https://www.hackthefuture.com/greenwood; Proof News and The Science, Technology, and Social Values Lab at the Institute for Advanced Study, "The AI Democracy Projects."

20 Carl Bybee defines "political epistemology" as "the politics of what we know and how we act as citizens is linked to the politics of how we know." See: Bybee, Carl. "Can Democracy Survive in the Post-Factual Age?: A Return to the Lippmann-Dewey Debate about the Politics of News." *Journalism & Communication Monographs* 1, no. 1 (March 1, 1999): 28–66. https://doi.org/10.1177/152263799900100103.

21 Lippmann, Walter. *Public Opinion*. Transaction Publishers, 2004, 10.

22 Warner, Michael. *Publics and Counterpublics*. New York: Zone Books, 2005, 84; Nancy Fraser, "Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy," *Social Text*, no. 25/26 (1990): 67, https://doi.org/10.2307/466240.

23 Dewey, *The Public and Its Problems: An Essay in Political Inquiry*, 161.

24 Dewey, 174.

25 Dewey cited in Matthew Festenstein, "Does Dewey Have an 'Epistemic Argument' for Democracy?," *Contemporary Pragmatism* 16, no. 2–3 (May 17, 2019): 219, https://doi.org/10.1163/18758185-01602005.

26 Dewey, *The Public and Its Problems: An Essay in Political Inquiry*, 232–33.

27 Melvin L. Rogers, "Revisiting The Public and Its Problems," in *The Public and Its Problems: An Essay in Political Inquiry* (Athens, Ohio: Swallow Press, 2016), 33.

28 Emily, interviewed on 12 October 2023.

29 For a more detailed account of reimagining legacy security frameworks to address the novel security threats and vulnerabilities that emerge with the rise of participatory technologies, specifically social media platforms, see Matt Goerzen, Elizabeth Anne Watkins, and Gabrielle Lim, "Entanglements and

Exploits: Sociotechnical Security as an Analytic Framework," 2019, https://www.usenix.org/conference/foci19/presentation/goerzen.

30 Sarah T. Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media*, Illustrated edition (New Haven: Yale University Press, 2019).

31 Goerzen, Watkins, and Lim, "Entanglements and Exploits."

32 Sam, interviewed on 29 September, 2023.

33 Stevie Bergman et al., "STELA: A Community-Centred Approach to Norm Elicitation for AI Alignment," *Scientific Reports* 14, no. 1 (March 19, 2024): 6616, https://doi.org/10.1038/s41598-024-56648-4.

34 Bergman et al., "STELA"; Afsaneh Rigot, "Design from the Margins," *Harvard Belfer Center for Science and International Affairs*, 2022, https://www.belfercenter.org/sites/default/files/files/publication/TAPP-Afsaneh_Design%20From%20the%20Margins_Final_220514.pdf; Christina N. Harrington, "The Forgotten Margins: What Is Community-Based Participatory Health Design Telling Us?," *Interactions* 27, no. 3 (April 17, 2020): 24–29, https://doi.org/10.1145/3386381.

35 Yuntao Bai et al., "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback" (arXiv, April 12, 2022), http://arxiv.org/abs/2204.05862; Anthropic, "Claude's Constitution," *Anthropic Blog* (blog), May 9, 2023, https://www.anthropic.com/news/claudes-constitution; Anthropic, "Collective Constitutional AI: Aligning a Language Model with Public Input," *Anthropic Blog* (blog), October 17, 2023, https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input.

36 Anthropic's constitutional AI is similar to polling and citizen assemblies proposed by other researchers to accomplish the same task. See, for example, Collective Intelligence Project, "The Collective Intelligence Project Whitepaper" (Collective Intelligence Project, 2023), https://cip.org/whitepaper; Aviv Ovadya, "'Generative CI' through Collective Response Systems" (arXiv, February 1, 2023), https://doi.org/10.48550/arXiv.2302.00672; Elizabeth Seger et al., "Democratising AI: Multiple Meanings, Goals, and Methods," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23 (New York, NY, USA: Association for Computing Machinery, 2023), 715–22, https://doi.org/10.1145/3600211.3604693.

37 Tyna Eloundou and Teddy Lee, "Democratic Inputs to AI Grant Program: Lessons Learned and Implementation Plans," *OpenAI* (blog), January 16, 2024, https://openai.com/blog/democratic-inputs-to-ai-grant-program-update.

38 Arvind Narayanan @random_walker, "Traditionally in ML, Building Models Is the Central Activity and Evaluation Is a Bit of an Afterthought. But the Story of ML over the Last Decade Is That...," *X.Com*, September 8, 2024, https://x.com/random_walker/status/1840731490239340896.

39 Kabir, interviewed on 17 November 2023.

40 Nanna Inie, Jonathan Stray, and Leon Derczynski, "Summon a Demon and Bind It: A Grounded Theory of LLM Red Teaming in the Wild" (arXiv, November 13, 2023), https://doi.org/10.48550/arXiv.2311.06237.

41 Shayne Longpre et al., "Position: A Safe Harbor for AI Evaluation and Red Teaming," in *Proceedings of the 41st International Conference on Machine Learning*, ed. Ruslan Salakhutdinov et al., vol. 235, Proceedings of Machine Learning Research (PMLR, 2024), 32691–710, https://proceedings.mlr.press/v235/longpre24a.html.

42 Sven Cattell, Avijit Ghosh, and Lucie-Aimée Kaffee, "Coordinated Flaw Disclosure for AI: Beyond Security Vulnerabilities," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, 2024, 267–80, https://doi.org/10.1609/aies.v7i1.31635; Allen D Householder et al., "The CERT Guide to Coordinated Vulnerability Disclosure," Special Report (CMU/SEI-2017-SR-022 CERT Division, August 2017), https://resources.sei.cmu.edu/asset_files/specialreport/2017_003_001_503340.pdf; Jonathan M. Spring et al., "On Managing Vulnerabilities in AI/ML Systems," in *New Security Paradigms Workshop 2020* (NSPW '20: New Security Paradigms Workshop 2020, Online USA: ACM, 2020), 111–26, https://doi.org/10.1145/3442167.3442177.

43 Seth Lazar, "Power and AI: Nature and Justification," in *The Oxford Handbook of AI Governance*, ed. Justin Bullock et al. (Oxford University Press, 2022), https://doi.org/10.1093/oxfordhb/9780197579329.013.12.

44 Burrell, Jenna, and Jacob Metcalf. "Introduction for the Special Issue."

45 United Nations and AI Advisory Body, "Governing AI for Humanity" (United Nations, September 2024), 42, https://www.un.org/en/ai-advisory-body; Rumman Chowdhury, "What the Global AI Governance Conversation Misses," *Foreign Policy*, September 19, 2024, https://foreignpolicy.com/2024/09/19/ai-governance-safety-global-majority-internet-access-regulation/.

46 We are aware that both within and outside of AI governance, extensive efforts have been made to improve the ability of safety and risk management paradigms on this front. We maintain that these improvements are necessary but not sufficient: without public interrogation, diffuse and contested social problems cannot be adequately addressed.

47 Melissa Heikkilä, "This New Data Poisoning Tool Lets Artists Fight Back against Generative AI," *MIT Technology Review*, October 23, 2023, https://www.technologyreview.com/2023/10/23/1082189/data-poisoning-artists-fight-generative-ai/; Trystan S. Goetze, "AI Art Is Theft: Labour, Extraction, and Exploitation: Or, On the Dangers of Stochastic Pollocks," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24 (New York, NY, USA: Association for Computing Machinery, 2024), 4, https://doi.org/10.1145/3630106.3658898.

48 Samantha, interviewed on 2 August 2023.

49 Hu and Singh, "Enrolling Citizens."

50 Nanna Inie, Jonathan Stray, and Leon Derczynski, "Summon a Demon and Bind It," 28.

51 Nico Grant, "Google Chatbot's A.I. Images Put People of Color in Nazi-Era Uniforms," *The New York Times*, February 22, 2024, sec. Technology, https://www.nytimes.com/2024/02/22/technology/google-gemini-german-uniforms.html.

52 Amari, interviewed on 6 October 2023.

53 Samantha, interviewed on 2 August 2023.

54 Zuri, interviewed on 27 September 2023.

55 Amari, interviewed on 6 October 2023.

56 Amari, interviewed on 6 October 2023.

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

58 Hu and Singh, "Enrolling Citizens."

59 Anthropic, "Challenges in Red Teaming AI Systems," June 12, 2024, https://www.anthropic.com/news/challenges-in-red-teaming-ai-systems.

60 See, Appendix #1 on design choices for genAI red-teaming in Singh et al., "Red-Teaming in the Public Interest."

61 Seth Lazar, "Legitimacy, Authority, and Democratic Duties of Explanation" (arXiv, October 11, 2023), http://arxiv.org/abs/2208.08628.

62 Reuben Binns, "Algorithmic Accountability and Public Reason," *Philosophy & Technology* 31, no. 4 (December 2018), https://doi.org/10.1007/s13347-017-0263-5; see also, Deven R Desai and Joshua A Kroll, "Trust but Verify: A Guide to Algorithms and the Law," *Harv. JL & Tech.* 31 (2017): 9; Victor Ojewale et al., "Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling" (arXiv, March 14, 2024), 2, http://arxiv.org/abs/2402.17861.

63 Reuben Binns, "Algorithmic Accountability and Public Reason," 545.

64 Nancy Leveson, *An Introduction to System Safety Engineering* (Cambridge, MA: The MIT Press, 2023), 50.

65 Sidney Dekker, *Drift into Failure: From Hunting Broken Components to Understanding Complex Systems* (Farnham ; Burlington, VT: Ashgate Pub, 2011), xii.

66 John Dewey, "(1939) Creative Democracy—The Task before Us," in *The Pragmatism Reader: From Peirce through the Present*, ed. Robert B. Talisse and Scott F. Aikin (Princeton, NJ Oxford: Princeton University Press, 2011), 151.

67 Dekker, *Drift into Failure*, 173; Jane Mansbridge, "Feminism and Democracy - The American Prospect," *The American Prospect*, February 19, 1990, https://prospect.org/civil-rights/feminism-democracy/; Fraser, "Rethinking the Public Sphere," 67; Elizabeth Anderson, "The Epistemology of Democracy," *Episteme* 3, no. 1–2 (June 2006): 12, https://doi.org/10.3366/epi.2006.3.1-2.8.

JACOB METCALF leads Data & Society's AI on the Ground program.

RANJIT SINGH is a senior researcher with the AI on the Ground team at the Data & Society (D&S) research institute.

BORHANE BLILI-HAMELIN is an ethicist at the AI Risk and Vulnerability Alliance (ARVA).

FILED UNDER **ESSAYS AND SCHOLARSHIP**