Aggregated Individual Reporting for Post-Deployment Evaluation

Jessica Dai*[1], Inioluwa Deborah Raji[1], Benjamin Recht[1], and Irene Y. Chen[1,2]

[1]*University of California, Berkeley*
[2]*University of California, San Francisco*

June 2025

Abstract

The need for developing model evaluations beyond static benchmarking, especially in the post-deployment phase, is now well-understood. At the same time, concerns about the concentration of power in deployed AI systems have sparked a keen interest in "democratic" or "public" AI. In this work, we bring these two ideas together by proposing mechanisms for *aggregated individual reporting* (AIR), a framework for post-deployment evaluation that relies on individual reports from the public. An AIR mechanism allow those who interact with a specific, deployed (AI) system to report when they feel that they may have experienced something problematic; these reports are then aggregated over time, with the goal of evaluating the relevant system in a fine-grained manner. This position paper argues that **individual experiences should be understood as an integral part of post-deployment evaluation, and that the scope of our proposed aggregated individual reporting mechanism is a practical path to that end.** On the one hand, individual reporting can identify substantively novel insights about safety and performance; on the other, aggregation can be uniquely useful for informing action. From a normative perspective, the post-deployment phase completes a missing piece in the conversation about "democratic" AI. As a pathway to implementation, we provide a workflow of concrete design decisions and pointers to areas requiring further research and methodological development.

*Correspondence to jessicadai@berkeley.edu.

1 Introduction

In the third week of April 2025, OpenAI quietly pushed an update to GPT-4o, the model that powers the ChatGPT product by default. Online, the complaints started rolling in: The newest update was annoying and introduced bugs; it overpromised and underdelivered. More frighteningly, it encouraged users to stop taking their medications, validated conspiracy theories, and worse. By April 29, around one week later, OpenAI had rolled back the change [OpenAI, 2025b].

In this case, the unstructured feedback of individual users was essential. OpenAI was unable to identify ahead of time that this "personality" update was problematic—in part because it is hard to anticipate the richness of usage patterns, and therefore failure modes. Users thought the problems were egregious enough that it motivated them to share online; enough users tweeted about the *same* problem that OpenAI noticed. This ChatGPT personality problem was serious, widespread, and was therefore caught even in the absence of a formal system to collect feedback. But what other, subtler, non-Twitter-viral patterns might be happening—and how might we find out?

In this work, we formalize *aggregated individual reporting* (AIR) as a general framework for understanding the real-world impact of an AI system in a structured, intentional, and thorough manner.[1] At a high level, an AIR allows those who interact with a specific AI system to submit feedback (reports) when they believe they have experienced harm due to the system. The mechanism aggregates these reports over time, with the goal of building collective knowledge about the contours of system behavior. Intuitively, one person having one bad experience does not by itself necessarily imply a system-level problem. On the other hand, if many reports of similar experiences begin to accumulate, then perhaps there may be an important underlying issue that requires action.

Our framework relies on two crucial beliefs: first, that those interacting with the system have unique and valuable perspectives on its failure modes, and second, that they ought to be able to express those perspectives in a nontrivial way. **This position paper argues that post-deployment evaluation must account for individual experiences—and that *aggregated individual reporting mechanisms*, as we define in this work, are a practical pathway to doing so.**

We do not claim to originate the concept of individual reporting. In fact, we are inspired by the existence of similar reporting mechanisms in other domains, like the Vaccine Adverse Events Reporting System (VAERS) [Shimabukuro et al., 2015], the Aviation Safety Reporting System (ASRS) [Beaubien and Baker, 2002], and various medical reporting systems [Wu et al., 2002]. We are also heavily influenced by calls to incorporate end-user expertise in algorithm evaluations (e.g., Shen et al. [2021], DeVos et al. [2022], Lam et al. [2022]) and to shift power towards the public (e.g., Kalluri [2020], Feffer et al. [2023]). Moreover, while several recent policy directives mandate the consideration of public feedback with specific reference to post-deployment monitoring (e.g., the EU AI Act [European Parliament, 2024], the UN General Assembly's first AI resolution [Assembly, 2024], and Biden's now-repealed executive order [Biden, 2023]), they are, by nature, only high level. By presenting a high-level structure for AIRs, we aim to offer a shared language as starting point for conversation across subfields and domains, and to concretize what effective implementation of this policy might look like.

The remainder of this paper is organized as follows. In Section 2, we define our idealized version of an AIR mechanism, and contextualize our definition with the current state-of-the-art for post-deployment evaluation and auditing for AI systems. In Section 3, we argue that individual reporting effectively identifies "unknown unknowns," and that aggregation enables downstream action. In Section 4, we elaborate on the normative case for individual reporting, placing our proposal in the context of recent calls for "democratic" AI. Section 5 describes more granular design decisions as well as associated open research questions. We conclude by discussing of key challenges for successful real-world deployment in Section 6.

[1] Throughout this paper, we will often use "AIR" as a stand-in for "aggregated individual reporting mechanism" for concision and clarity, and to distinguish the full mechanism from its components (aggregation and individual reporting).

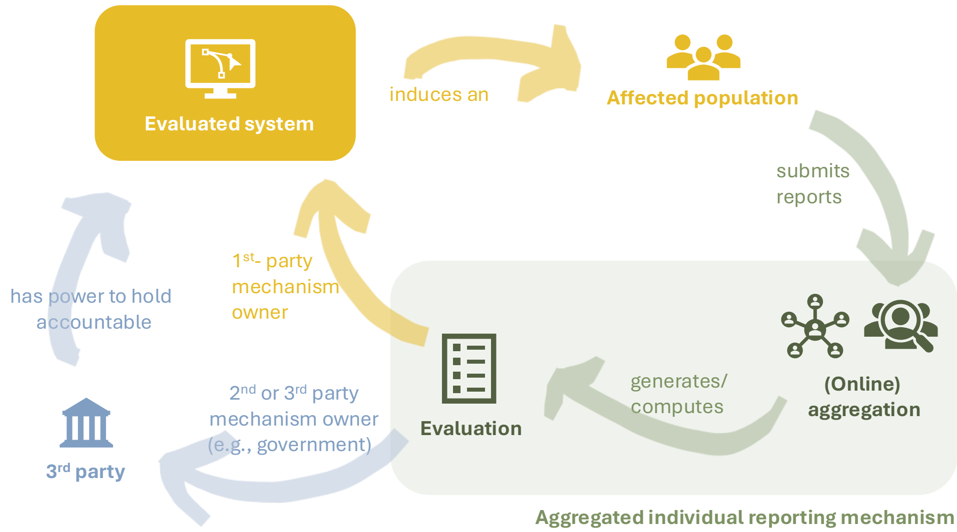

Figure 1: The components of our framework and how they interact.

2 Defining aggregated individual reporting

We begin by describing our idealized vision of a mechanism for aggregated individual reporting, illustrated in Figure 1. This definition is intentionally broad, so as to cover a wide range of potential applications—see Section 5 for more granular details that describe various potential implementations. Even with this broad scope, we note that, to the best of our knowledge, (public or non-proprietary) AIRs for AI systems are not currently in mainstream use.

2.1 Defining an idealized individual reporting mechanism

Our understanding of an AIR mechanism begins with three key entities.

First, the ***evaluated system*** determines the scope of the AIR mechanism. For example, this could be a general-purpose model like ChatGPT or Claude; a product that scribes in-person doctors' appointments; or a predictive algorithm that several banks use for making loan decisions. Reports are submitted *about* the evaluated system. Second, the evaluated system induces an ***affected population***. This could include users of ChatGPT, or their loved ones; patients and healthcare workers; or loan applicants. Reports are submitted *by* members of the affected population. Finally, the AIR mechanism is orchestrated by a ***mechanism administrator***, which is the organizational entity that collects and aggregates the reports. The administrator makes key decisions like determining the scope of the evaluated system and the affected population, as well as various implementation details as discussed in Section 5. The mechanism administrator could be first-party, i.e. the same organization that operates the evaluated system (e.g., OpenAI collecting reports for ChatGPT); second-party, i.e. an organizational user of the evaluated system (e.g., a hospital system collecting reports about an AI scribe product); or third-party, i.e. an external organization (e.g., a government body or activist nonprofit collecting reports about a loan allocation algorithm). This taxonomy matches prior work in the AI audit space (e.g., Raji et al. [2022]).

AIR mechanisms comprise the following components involving these entities.

(1) *Individual reporting*: Reports are submitted by members of the affected population about specific experiences with the evaluated system.

(2) *Aggregation for evaluation*: Reports are aggregated and interpreted over time. The goal of such aggregation is evaluation: to understand or describe the behavior of the evaluated system in a fine-grained manner.

(3) *Evaluation-conditional action:* The aggregated evaluation supports downstream action. That is, there are evaluation outcomes where, if and when the reports are consistent with those outcomes, the mechanism administrator takes associated action.

Ensuring that reports are tied to the *evaluated system* makes it possible for the aggregation to generate specific insights about system behavior; this specificity allows for downstream action, in that it explicitly describes some (undesirable) property of the system that has emerged. The temporal component is critical—continuous evaluations make it possible to understand changing use cases and experiences over time, and moreover, to identify problems (and take action) quickly as they arise.

In Figure 2, we give four examples of AIR mechanisms. In the first row, we show how VAERS, an existing reporting system, implements the criteria of our framework. The second row presents a hypothetical problem setting following an example in Dai et al. [2025], and the third and fourth rows highlight speculative examples for applications of AIR mechanisms. As these examples indicate, AIRs can be established for various systems (in fact, the evaluated system need not be algorithmic at all); moreover, the implementation of the mechanism depends closely on the type of harm that an AIR is designed to surface.

Evaluated system	Mechanism administrator	Affected population	Individual report information	Evaluation condition (when would downstream action be taken due to patterns in reports?)	Downstream actions
FDA-approved vaccines (deployed in real-world system as VAERS)	3rd-party (United States CDC & FDA)	All patients who received a particular vaccine	Specific vaccine (brand and batch), specific adverse event (e.g.,) and demographic information	Elevated overall frequency of adverse event reports compared to expected baseline frequency e.g., *myocarditis appears frequently for the COVID-19 vaccine.*	Further investigation of specific vaccine-side effect pairs (e.g. additional research or data collection), and notification of relevant parties (e.g. published reports)
Loan allocation algorithm at Bank X (hypothetical from Dai et al. 2025)	3rd-party (activist organization)	All loan applicants to Bank X	Demographic information and the claim of potential discrimination	Identification of a subgroup that experiences disproportionate rates of harm e.g., *financially-healthy Black applicants are denied loans at a higher rate.*	Gathering evidence to initiate a legal discrimination case
AI medical scribe product (speculative example)	2nd-party (hospital system client)	Healthcare workers who use the tool, and their patients	Free-text notes and information about reporter; scribe text and edit history, for healthcare worker reporters	Clinically-relevant failure modes of the scribe product. e.g., *AI scribe repeatedly makes errors for visits about pregnancy complications.*	Feedback provided to company developing AI scribe; temporary usage guidelines given to clinicians working in (e.g.) maternal health
ChatGPT (speculative example)	1st-party (OpenAI)	All users of the ChatGPT product	Chat transcript and free-text notes	Wide-scale safety-critical behavior. e.g., *the newest model exhibits dangerously sycophantic behavior.*	Rollback to prior model version; post-mortem conducted for flaws in internal evaluations.

Figure 2: *Examples of how AIRs could be set up for a variety of applications. Here, we elide the corresponding aggregation methods in order to focus on the application itself; note, however, that the "evaluation conditions" are aggregate system failures rather than per-report problems.*

2.2 Current state-of-the-art in post-deployment evaluation

The idealized definition given in Section 2.1 excludes most prominent systems that currently exist for crowdsourcing and post-deployment monitoring for AI. Generally speaking, existing systems appear to fall into three categories: third-party collections of real-world problems with AI systems; general approaches to post-deployment evaluation; and targeted flaw disclosure approaches like bug bounties and red-teaming. We briefly clarify the relationship of aggregated individual reporting mechanisms to these approaches.

In the first category are several well-known *incident databases and risk repositories*, which include the website `incidentdatabase.ai`, where incident submissions are available to the public [McGregor, 2021]; the OECD's AI Incidents and hazards Monitor (AIM), which is an automated system that scrapes all AI-related news headlines globally [OECD, 2023]; and the MIT AI Risk Repository, which tags individual incidents from `incidentdatabase.ai` with additional metadata from the Risk Repository framework [Slattery et al., 2024] and is available to download as a Google sheet. While the exact implementation of each database is slightly different, the high-level commonality is that they tend to collect discrete, externally-submitted *incidents*. These incidents are typically news events that were related in some way to *any* deployed AI system, with a general focus on examples of misuse.[2] As a result, these collections of incidents are much broader than individual interactions that can become descriptive of model behavior at a more granular level; in fact, aggregation of these incidents happens only to the extent of tabulating the approximate frequency of these discrete news events across impact category and severity.

The second category includes most current approaches to *post-deployment monitoring* or evaluation; here, we highlight the approaches that are most closely related. For understanding real-world usage of Claude, Anthropic has developed a system called Clio, which applies k-means clustering to a fixed batch of chat transcripts, then post-hoc labels the clusters with Claude [Tamkin et al., 2024]. Unlike the incident databases described above, Clio is specific to an evaluated system (Claude)—and, notably, the goal of Clio is specifically to identify "unknown unknowns" in usage patterns. However, the collection of all chats is quite different from individually-submitted reports that highlight problematic behavior, and thus Clio captures different insights than an AIR would. Moreover, at least according to public information, Clio operates on static batches of transcripts, so it is unknown how or whether it can identify trends that develop over time. For large language models in general, Chatbot Arena has become a popular evaluation platform for ranking multiple LLMs [Chiang et al., 2024]. While Chatbot Arena does rely on real-time, crowdsourced feedback, the goal of their evaluations is primarily to compare LLMs (by generating a ranking), rather than conducting fine-grained evaluations that could later inform downstream action; moreover, the form of feedback afforded by the platform is rather limited as a "report."

Finally, we note that *flaw disclosure* mechanisms, e.g., as outlined in Longpre et al. [2025], are an important starting point, but are insufficient for our goals. That is, an aggregated individual reporting mechanism can—and likely should—include much of the flaw disclosure machinery outlined in Longpre et al. [2025], but a flaw disclosure system by itself would not necessarily qualify as an AIR. In the context of our definition, flaw disclosure systems may sometimes satisfy the first condition (*individual reporting*), but do not include the other criteria (aggregation for evaluation and downstream action). Thus, our specific proposal is explicitly narrower.[3] This can be seen when considering concrete instantiations of flaw disclosure mechanisms, such as bug bounties

[2] For instance, at time of writing, the most recent incident at `incidentdatabase.ai` is that the New Orleans police department used banned facial-recognition software (Incident 1075). The harm being described in this incident is about NOPD's usage of a particular system, rather than a specific instance or individual experience of that system's failure.

[3] Regarding the specific manuscript presented in Longpre et al. [2025], the contribution of our work is distinct, but complementary. We are explicitly focused on public reporting and understanding its benefits, as in Sections 3 and 4 (whereas non-expert reports are are only briefly mentioned in their work); we also emphasize methodological problems and directions for the AI research community (Section 5). We refer readers to their manuscript for a thorough legal and policy overview and a taxonomy of current flaw disclosure methods, which our manuscript excludes.

for algorithmic problems and red-teaming. The former is typically focused on resolving bugs on a per-report level, rather than understanding problems and taking action on an aggregate level (e.g., as discussed in Kenway et al. [2022], with the bias-bounty mechanism of Globus-Harris et al. [2022] as a rare exception); the latter often focuses on evaluations at the pre-deployment phase, and solicits participation from predefined groups rather than all members of the affected population (e.g., Ahmad et al. [2025]).

We see our proposal as complementary to all of these approaches, which are useful and important parts of the AI evaluation ecosystem. AIRs, as we have defined them, enable distinct types of evaluations beyond what is already covered by these approaches—but cannot, by themselves, supersede these efforts.

3 Aggregated individual reporting enables actionable harm discovery

In this section, we argue that AIRs, as outlined in Section 2.1, have concrete benefits beyond what existing systems can provide. Prior work, especially in human-computer interaction, has documented a desire for ways to enable auditing and evaluation of algorithmic systems from a wider population. For example, user-driven auditing (e.g., Shen et al. [2021], DeVos et al. [2022], Deng et al. [2023], Lam et al. [2022]) has been studied as a means for eliciting user feedback that identifies problems with an algorithmic system beyond centralized evaluations. Shared concerns and challenges presented by these works, which have primarily involved small-scale case studies, include aggregating and interpreting feedback, and using feedback to drive downstream action. In fact, Ojewale et al. [2025] explicitly call for the development of "tools for harm discovery" and "participatory methods" as a pathway towards accountability.

Our framework seeks to formalize some these intuitions, and each component of our definition in 2.1 plays a specific role: individual reports enable harm discovery by surfacing unknown unknowns, while aggregation is a prerequisite to making them actionable.

3.1 Individual reports surface unknown unknowns

In domains where reporting systems are well-established, it is widely understood that reports contain useful information that may otherwise never be identified (e.g., see Beaubien and Baker [2002] for a decades-old review in aviation, and Wu et al. [2002] for health). In our vision of individual reporting, however, the affected population is a much wider set of people, beyond domain experts and practitioners. What kinds of insights might we expect to see from their reports?

Prior work suggests that audit-style feedback from "end users" draws from their prior beliefs and experiences, and can reveal clusters of shared problems (including some that were previously unknown or unaccounted for), as well as identify meaningful disagreement across users [Lam et al., 2022, DeVos et al., 2022]. For concreteness, we will focus here on social media platforms like Twitter/X as as case studies for (informal) individual reporting.

The Twitter app is of course not a purpose-built reporting system, and there is therefore no aggregation mechanism dedicated to explicitly making sense of report content. However, individuals frequently take to social media to share their experiences with specific algorithms. Shen et al. [2021] term this phenomenon "everyday algorithm auditing," and show how social media platforms enable an organic, informal process for both raising awareness and validating problems raised by initial reports. In their analysis, they highlight that "everyday audits" leverage the lived experience (e.g., cultural background) and situated knowledge (e.g., application contexts) of everyday users.

Returning to the motivating example at the beginning of this paper, the analysis of Shen et al. [2021] is largely consistent with the patterns that can be seen from posts about ChatGPT sycophancy. For example, users noted performance degradation even in quantitative tasks [Ho, 2025, Laura, 2025], reflecting situated knowledge with respect to users' expectations in specific domain applications; more serious posts highlighted scenarios where ChatGPT validated flat-earth and conspiracy beliefs [fortheloveoftheworld, 2025], reflecting both lived experience and situated knowledge.

A subtly distinct pattern from that arose in reports about sycophancy, compared to trends identified by Shen et al. [2021], is reports that reflect user creativity—rather than realistic practical usage—in a way that can be thought of as informal "red teaming." For example, users showed example output responses when given prompts about math and "pickle rick" [Williawa, 2025], monologues by unsavory fictional characters [Bharath, 2025], as well as genuinely-safety critical responses elicited intentionally [Reviews, 2025, Frye, 2025].

Importantly, the sycophancy case study also reflects a limitation of relying on social media as a vehicle for "reporting." The *types* of problems that can be identified via these "reports," somewhat by definition, are those that are amenable to virality. In fact, the methodology of the taxonomy in Shen et al. [2021] also relied on prior knowledge of cases that were "high-profile" on social media. Sycophancy is a prime example of a virality-friendly problem: It appeared to affect all users across use cases and backgrounds, so that various users felt empowered to share their particular experience that reflected the underlying problem. Moreover, it was easy to produce content that illustrated sycophancy, but was also (e.g.) highly humorous or inflammatory. However, many serious model flaws are not so "clickworthy"; some failures might only affect a small slice of users, and in ways that cannot be constructed as a popular tweet. More systematic approaches, therefore, are necessary.

3.2 Aggregation for actionability and accountability

We now argue that, beyond individual reporting, *aggregation* is necessary to make use of the information contained in reports. Not all potential problems with an evaluated system are inherently statistical; in many cases, however, individual experiences can only be understood in the context of aggregated evaluations. For example, if the goal is to identify subgroups that disproportionately experience harm, as in Dai et al. [2025], then individual reports become significant insofar as they can be described statistically; similarly, participants in DeVos et al. [2022] study mention uncertainty about whether particular algorithmic behaviors ought to be considered examples of harm, especially in the absence of knowledge about how others experienced the system.

For areas where report databases (often called "incident databases") are already established as a component of safety monitoring, it is well-understood that that focusing on rectifying individual incidents is limited. Instead, the goal should be instead to find high-level patterns that can inform future procedural changes; in these fields, learning from reports has indeed been effective for shaping future practice [Macrae, 2016, Jacobsson et al., 2012, Robinson, 2019].

Despite this, many reporting systems do not include intentional aggregation as a core component; these systems appear to have had limited impact beyond the scope of reports themselves. On the other hand, the few examples of successful action from crowdsourced information have involved aggregation, suggesting that aggregation is a necessary (though potentially not sufficient) condition for taking high-level action from reports. In the remainder of this section, we discuss several case studies of crowdsourced or public reporting feedback, and the extent to which they were successfully spurred concrete action.

CFPB Consumer Complaints, VAERS, and different levels of actionability. The U.S. Consumer Finance Protection Bureau (CFPB) maintains a Consumer Complaint Database to which members of the public are eligible to submit. The CFPB itself does not directly analyze or address these complaints Consumer Financial Protection Bureau [2012]; instead, it acts as an intermediary, passing the complaints to the relevant financial institutions and mandating direct responses from the financial institution [Littwin, 2015]. Though the complaints submitted by individuals do actually receive responses from the responsible parties (i.e. banks), the emphasis is on directly addressing individual incidents, rather than the problems that emerge when considering them all collectively. In this way, the Consumer Complaint Database resembles the incident databases described in Section 2.2 and the flaw disclosure framework from Section 2.1.

To the best of our knowledge, insights from the CFPB complaint database as a whole has not triggered further enforcement or legislative action; this is perhaps unsurprising, given the emphasis on actionability for individual complaints, rather than at in aggregate. Several academic works have found problematic trends by analyzing complaints collectively (e.g., Bastani et al. [2019], Ayres et al. [2013], Haendler and Heimer [2021]); thus, this focus on per-complaint resolution is a design choice, not an inherent limitation of the data itself.

On the other hand, the VAERS database is continually monitored explicitly *for* aggregate-level harm. For VAERS, the event of concern is not any individual report of an adverse event. Rather, the concern is whether reports for particular vaccines occur with abnormally high frequency; examples of clinically-relevant side effects initially discovered via VAERS abound [Shimabukuro et al., 2015, Singleton et al., 1999].

Beyond initial identification, another important usage of VAERS has been post-hoc investigation of problems that were originally flagged via case study. For instance, it is now well-known that the COVID-19 vaccines induced an elevated risk of myocarditis, but only in younger men; however, in the early stages of vaccine rollout, the conversation was limited to various healthcare providers noticing, via case study, that myocarditis appeared to be a common occurrence overall [Mouch et al., 2021, Larson et al., 2021, Marshall et al., 2021]. A more holistic understanding of the issue, including that myocarditis appeared to be limited to younger men, was attained only after post-hoc analysis of VAERS reports [Witberg et al., 2021, Oster et al., 2022]. This, in fact, was one motivating example for the method proposed in Dai et al. [2025]: if VAERS reports about myocarditis had been analyzed with a disaggregation over demographic identity, then VAERS reports could have directly confirmed the affected subgroup more quickly.

ChatGPT sycophancy, "spiritual delusions", and the limitations of informal aggregation. We conjecture that one reason that the ChatGPT sycophancy problem resulted in decisive change was that the Twitter timeline algorithm, and the platform's affordances of likes and retweets, served as a quasi-aggregation scheme. The brief dominance of tweets highlighting sycophancy on the timeline showed that this was a problem that impacted a wide range of users, and that there was general agreement that the behavior was problematic. In a move that is remarkably unique for tech companies, this culminated in explicit action by OpenAI to roll back the new model deployment, and to initiate some discussion of what went wrong [OpenAI, 2025b,a]. In some sense, it was "lucky" that this particular safety problem happened to have been friendly for virality; in general, there is no guarantee that quasi-reports to social media are necessarily seen, or taken seriously, by those who have control over the evaluated system.

It is instructive, here, to make a comparison to some approximately-contemporaneous Reddit threads, which detailed severe mental health crises due to what appear to be ChatGPT-caused "spiritual delusions" (first referenced in Klee [2025]; the more recent Hill [2025] describes similar phenomenons). Notably, many cases mentioned had been ongoing for weeks, and therefore could not be entirely attributed to the late-April update. There has been no blog post that explicitly addresses these problems—and, indeed, no reason to think one might be forthcoming, based on the company's statements in the reported articles.

As for why this might be the case, the obvious reason is that the "spiritual delusion" problem is likely more complex than the sudden increase in sycophancy, which was easily addressed by a rollback. However, we speculate that perhaps one additional factor in the lack of decisive, publicized action from OpenAI is that, overall, the "aggregation mechanism" of Reddit was much less powerful than on Twitter. The conversation about long-run psychological impact did surface beyond Reddit, where the posts were originally made, to reported features in major national magazines. However, both Reddit as a platform and reported stories as a format emphasize individual-level narratives, which are less compelling as indictments of systematic behavior patterns.

Targeted crowdsourcing and the value of statistical evidence. Finally, we discuss two systems that crowdsourced targeted surveys of specific algorithmic systems. While the scope of these surveys were narrower—meant to discover average rates of pre-specificed metrics, rather than general evaluation—they are notable because their aggregations provided concrete statistical evidence for individual experiences.

In 2020 and 2021, Mozilla led a study using a browser extension called *RegretsReporter*, which crowdsourced information about the experience of Youtube recommendations [McCrosky and Geurkink, 2021]. The study found that that recommendations from the Youtube algorithm were disproportionately responsible for serving content that users regretted seeing (and violated terms of service)—and moreover, that non-English speakers were most seriously affected.[4] While Youtube never publicly disclosed whether specific algorithmic changes were made in response (a weakness we discuss in Section 6), the company did directly address the report in public statements [Klar, 2021, Lawler, 2021, The Next Web, 2021].

Perhaps more optimistically, *Fairfare* is a system that crowdsources information on rideshare wages, with the goal of understanding the extent to which drivers were being underpaid as well as overall average pay rates [Calacci et al., 2025]. This information not only empowered drivers to organize, but also led to legislative impact that was motivated by statistics computed via Fairfare.

More broadly, statistical evidence—in contrast to, e.g., anecdotal accounts—is especially useful as a means for pressuring *organizations* or *institutions* to make change [Recht, 2025]. For example, business leaders are more amenable to making decisions that appear "data-driven," rather than responsive to anecdotal experiences [H. Davenport, 2014]. Statistical evidence is also treated differently in litigation contexts (see, e.g., Espeland and Vannebo [2007]). The question of what *kinds* of statistical evidence are considered acceptable will depend on the particular application context, and not all harms are inherently statistical. At the very least, however, the language of statistics can empower individuals by validating their experiences. Both RegretsReporter and Fairfare were simply *formalizing* folk intuitions that Youtube users and rideshare drivers individually already had—but the collection and aggregation of data made it impossible to dismiss those perspectives as individual experiences of one-off anomalies.

4 Aggregated individual reporting as a pathway towards "democratic" AI

In this section, we take a brief detour from practical benefits to discuss the normative underpinnings of our proposal.[5] In recent years, the recognition that modern AI systems both require and accelerate the concentration of power has spurred a flurry of research on how AI systems might be designed through quasi-democratic processes—"participatory" [Birhane et al., 2022, Delgado et al., 2023, Gilman, 2023], "pluralistic" [Dai and Fleisig, 2024, Sorensen et al., 2024a,b], "collective" [Huang et al., 2024], and so on. Methodological research in these directions focus almost exclusively on the development phase, and rarely consider how "the public" might engage with AI systems

[4] The latter is also an example of crowdsourced data finding new, non-obvious insights, as discussed in 3.1.

[5] While of course the normative and practical cannot be fully disentangled—indeed, Estlund's core thesis about democracy is that it is practically effective exactly because it is normatively desirable, and vice versa [Estlund, 1997]—the arguments we present in this section are those that rely on conceptual foundations.

post-deployment—despite calls for increasing the power of "the public" from more conceptual work (e.g. Feffer et al. [2023]).

The "democracy" analogy relies on a rough analogy between AI systems and governmental bodies (about which democracy is classically theorized). For instance, these works commonly ask, how might "democratic" inputs shape the model spec or training objectives? Implicit in this question is an analogy to the same way that "democratic" inputs determine the outcome of an election.

We argue that such an analogy, while not incorrect, is incomplete. A key tenet of democracy is *consent of the governed*: the ability for members of the public to express their will about governance, and a process for those expressions of will to have meaningful bearing on future outcomes.[6] Crucially, such consent is not a singular event; instead, it is an ongoing process that, theoretically, must continue for as long as the governing body remains in power [Bertram, 2023, Gourevitch and Rousseau, 2018]. Democratic legitimacy derives not just from the fact that citizens can shape the parameters of their government's future actions, but also from their ability to provide input on ongoing action—to revoke consent.

A critical ingredient for a "democratic" approach to AI, therefore, is the ability for members of the public to collectively raise issues with systems *after* they have already been deployed. In the same way that a democratic governing body (a kratos) *requires* input from its citizens (demos) to continue effective governance, those who operate an AI system cannot fully understand the real-world behavior of their system without the input of those who have interacted with it. And, in the same way that the demos *ought* to have its concerns be heard by its kratos, those who interact with possibly-consequential systems also *deserve* for their concerns to be taken seriously, and systematically, by those who build the systems.

In other words, "democratic AI" is not just a design problem; it is an accountability problem. Our framework seeks to take a step towards this by not only offering the ability for individuals to provide feedback—via reporting—but also providing an avenue for them to be heard. As a starting point, aggregation is a lens through which (e.g.) the owners of the evaluated system can understand and interpret feedback ("helping the state see," in the sense of Scott [1998], by surfacing details that centralized evaluations are unable to understand). More theoretically, aggregated reports also have the potential to develop and shape the "general will," in the sense of Rousseau, from a collection of expressions of "individual wills" [Gourevitch and Rousseau, 2018].

From a more concrete perspective, one recent work that seeks to place AI governance mechanisms in the context of "democratic" processes is the Democracy Levels framework of Ovadya et al. [2024]. This framework rests on a conception of "democratic processes" as involving a remit (scope), constituent population, and an output decision; the hierarchy of "levels" corresponds to the extent to which the outcomes of this process have binding power over the AI system being governed.

Like the works mentioned above, the examples given in the Democracy Levels framework are primarily about pre-deployment rules and guidelines. However, we argue that mechanisms fo aggregated individual reporting can be seen as an almost direct stand-in their understanding of a "democratic process": the constituent population can be seen as our affected population, the remit can be thought of as our evaluated system, and the decision can be thought of the evaluation computed from the aggregated reports.[7] In fact, the mechanisms we propose can be direct drop-ins to their "levels"—for instance, in Level 0, aggregated reports have no impact; in Level 1, reports are accepted but do not necessarily affect the deployed system; in Level 2, the results of aggregated evaluation directly imply a default response action.

Finally, we note that, in political theory, the very notion of a "public" is a complex and contested notion. For instance, user-led audits have been proposed as pathways to developing *counterpublics*,

[6]This is, e.g., canonically expressed in Locke [1988].

[7]One interesting difference is that their definition of "democratic process" assumes a one-time event that can be initiated by various discrete triggers, whereas we think of our mechanism as continuous.

where individuals that are marginalized with respect to (or otherwise in opposition to) the wider "public" come together to collectively build knowledge and take action [Fraser, 1990, Baik and Sridharan, 2024, Shen et al., 2021, DeVos et al., 2022]. Our discussion in this section has not explicitliy considered the distinction between a hegemonic "public" as opposed to marginalized "counterpublics," instead generally emphasizing the value of giving voice to a *generic* public. We leave discussion of this distinction to future work; we expect that whether AIRs do empower counterpublics will depend on the details of how practical implementation plays out.

5 Towards implementation: practical details and open research questions

We now turn to discussing design decisions that must be made, either implicitly or explicitly, in the implementation of any AIR; each step also naturally gives rise to various multidisciplinary research questions. In Figures 3 and 4, we outline the relevant questions in detail.

5.1 Concrete design decisions

Here, we overview the high-level categories of design decisions and some implications of those choices. While these decisions are interdependent—reporting affordances also affect what aggregation methods would be effective, as well as what types of evaluations are available—there are a wide range of ways in which these decisions could be made, depending on context.

Within this section, we use Dai et al. [2025] as a running example of a methodological proposal that is largely consistent with our framework; this work proposes individual reporting for post-deployment fairness auditing, and provides algorithms that are specific to this task.

Organizational decisions. The first core set of decisions that must be made are organizational: what entity will take the role of *mechanism administrator*, and what relationship will it have to the *evaluated system*? This decision affects what kinds of problems the mechanism hopes to identify (i.e., the end-goal of the evaluation) as well as the nature of available downstream action.

A first-party administrator will have the fullest information about the evaluated system (e.g., when particular feature rollouts or updates were made), and be able to quickly gather additional data that may become relevant in order to contextualize information raised by reports. Since the first-party administrator has control over the evaluated system, the organization can also directly make changes in reponse to evaluation results. However, a first-party administrator may also inadvertently restrict the scope of the system, or intentionally choose to ignore the evaluation. On the other end of the spectrum, a third-party administrator would have nearly no additional information beyond what is included among reports, and cannot directly update or improve the system. However, such an organization could bring external leverage to the evaluated system, e.g. via media or legal pressure.

Dai et al. [2025] is developed under an assumption that the goal would be to identify the subgroups that are harmed, but, as a primarily methodological work, it does not specify what organizational entity would be the administrator, or what downstream action might look like.

Reporting affordances. What information is included in a report, and how do reporters experience the process of reporting? Examples of potential report formats can be seen in the fourth column of Figure 2, as well as the the flaw disclosure worksheet in Longpre et al. [2025]. For the algorithm proposed by Dai et al. [2025], the only information collected in a report is demographic information about the reporter; however, one could naturally imagine that reports could

Design decision	Options and examples	Example research questions
Organizational		
Mechanism administrator & relation to evaluated system	• First-party: same org (e.g. OpenAI for ChatGPT) • Second-party: user of evaluated system (e.g., hospital system for AI scribe product) • Third-party: External org (e.g., government or activist nonprofit)	• What arrangement is "optimal" or incentive-compatible? • How do individuals within these organizations conceive of pathways to impact?
Scope of "affected population"	• System users (e.g., ChatGPT users) • System users and those close to them (e.g. friends/family of ChatGPT users) • System users and non-users affected by system usage (e.g., healthcare workers and patients)	• How does the inclusion of different "user roles" affect the substance of report content?
Individual reporting (*Reports are submitted by members of the affected population about specific experiences with the evaluated system*)		
Publicizing the reporting mechanism	• Reporting option directly available at or after each system interaction (e.g., "submit report" available in UI, or sent as part of follow-up) • Advertisements on social media about the opportunity to submit reports	• What pathway is the most effective? • How do publicity methods affect who submits reports and why?
Handling reporting behavior	• Submitted reports are taken "as is" with the understanding that reporting • Assumptions are made about reporting behavior (e.g., that heterogeneity in reporting rates is not too extreme) • Side information (outside of submitted reports) is sought in order to characterize reporting behavior (e.g., choosing some subset of reports to "verify")	• Who is more likely to report, and why? What topics/types of problems are more likely to generate reports, and why? • How can we model and detect disparate rates of reporting for various reasons? How can we draw inferences that account for or are robust to uncertainties in reports?
Report content and affordances	• About specific one-off interactions (e.g., "I said X and the model responded Y") • About longer-run series of interactions (e.g., "over the last 2 weeks, the model has been telling my partner that they are a 'chosen one'") • May or may not contain sufficient "state" to completely reproduce the problem (e.g., real chat transcript may or may not be available) • May or may not contain additional contextual information about the reporter or impacted party (e.g., demographics, context on usage, location, timestamp, etc.) • May or may not include information about believed severity, justification, or proposed solution (e.g., "this was a minor problem that could worsen"; "this was problematic due to X"; "it would have been better if Y happened instead")	• How do different affordances in report format enable reports about different types of harm? • How can we quantitatively process different types of information, including rich unstructured data like free text, and incorporate them into a quantitative (possibly-statistical) aggregation?
Incentivizing and encouraging reports	• No additional incentive for reporters • Financial compensation for reports that turn out to reflect what is later deemed to be a 'true' problem • Report affordances are designed to shape reporting behavior (e.g., intermediate report summaries are made visible to the public, so that reporters may be interested in contributing to the conversation about currently-leading concerns)	• What motivates potential reporters to actually submit reports? • Are there concerns about misaligned incentives that might affect the 'quality' of reports? • Are there feedback loops that might arise?

Figure 3: *Organizational and interaction-focused questions.*

include more information depending on the application, such as medical history (for a vaccine or pharmaceutical system) or financial background information (for a loan allocation system).

Reporting behavior. Due to the nature of reporting data it is, intrinsically, essential to understand reporting behavior: what factors affect the decision to submit a report, and, crucially, in what ways might reports be correlated to the target of evaluation? Do different subpopulations report

Design decision	Options and examples	Example research questions
Online aggregation (*Reports are aggregated and interpreted over time; the goal is describing system behavior in a fine-grained manner*)		
Evaluation trigger (what evaluation seeks to identify, and what would trigger downstream action)	• Set up with specific type of harm in mind; try to identify the specific harm as soon as possible if it does arise • Set up with a more unstructured approach; try to identify important trends that emerge over time, and allow	• What methods can successfully achieve these goals in an online or quasi-online manner? How do they adapt to or handle changes over time? • What methods can adapt to richer "hypotheses"? • How can we balance batching data with handling true sequentiality?
Type of evidence as an output of aggregation	• A rigorous statistical guarantee is required (e.g., "with probability 1-α, the findings of the aggregation are significant…") • More ad-hoc interpretation is sufficient (e.g., looking at the output of a clustering algorithm)	• What are the relative strengths and weaknesses of various types of evidence for the purpose of motivating downstream action? • What are the tradeoffs involved in pursuing more vs. less "rigorous" outcomes? What methods strike a balance between data efficiency vs. validity of conclusion?
Mechanism for action (*Some evaluation results may trigger action; the mechanism administrator is responsible for doing so.*)		
Types of action that can be taken once evaluation trigger is reached	• First-party administrators: Rollback to prior model version (e.g., if new problems suddenly arise after new deployment) • Second-party administrators: Revisit usage policy (e.g., internal guidelines for when a tool should be used or not) and/or provide feedback to owner of evaluated system • Third-party administrators: Build public pressure (e.g., via media) and/or action towards accountability (e.g., legal case against first-party) • All administrators: may use the evaluation trigger as a starting point for further investigation (e.g., additional data collection)	• What kinds of evidence can compel action from an external third party? • How can closed systems be pressured to admit a reporting system?
Ongoing contact with affected population	• Public notice of problem identified & changes made • Specific outreach to reporters who noticed the problem	• To what extent does, and should, the reporting mechanism encourage long-term engagement from reporters?

Figure 4: *Methodology-focused questions.*

at different rates? Do different types of issues lead to different reporting behaviors? In Dai et al. [2025], the choice is to commit to a set of (quantitative) assumptions about the extent to which reporting rates can vary, and incorporate those assumptions into the design of the algorithm.

Aggregation method and evaluation condition. Given the affordances offered to reporters, what method will be ued to interpret them over time—that is, how are evaluation results computed? Given that method, what specific results would define the evaluation condition (i.e. trigger for downstream action)? Methods should be sequential, or at least explicitly consider a temporal component (e.g., via sequences of batched data). This is because AIRs should be accessible to reporters at any time (rather than within the scope of a centralized study with a defined start and end date), as well as the classic distribution shift problem: users' needs in relation to an evaluated system will change over time, as will the system itself. In Dai et al. [2025], the method is formalized as a sequential hypothesis test with type-I error control. Therefore, the evaluation result that triggers action is exactly when the "null hypothesis" of no overrepresented subgroups can be rejected at level α. A failed hypothesis test might, for instance, prompt further inquiry into the reported harm for a particular subgroup.

5.2 Active research communities

Finally, we highlight some examples of existing lines of work within the AI research community—beyond auditing and evaluation—that can effectively inform these design questions.[8] While these examples are non-exhaustive, we hope to illustrate that AIRs can benefit from a wide range of methodological and disciplinary perspectives.

Interaction design. From prior work, we know that the affordances that are available to users to share feedback affect the content they share (e.g., social media platforms and virality-friendly content; see also discussion in DeVos et al. [2022] about the importance of platform affordances). What kinds of designs can encourage the most effective reporting feedback (e.g., can facilitate long term engagement, as discussed in Deng et al. [2023])? Are checklists, as formalized in Longpre et al. [2025], sufficient? Should intermediate evaluation results be made public, and if so, how? E.g., if individuals see that others have reported similar problems, does that encourage them to submit their own reports?

Reporting rates and behavior. Empirical studies of existing reporting systems have shown that different subgroups often report different types of problems at different rates (e.g. Agostini et al. [2024], Liu and Garg [2022], Liu et al. [2024]); at the same time, these works also suggest the potential for statistical methods that can estimate or correct for varying reporting rates. As the crowdsourcing experiments of Globus-Harris et al. [2024] suggest, participants can occasionally be adversarial, and the details of mechanism implementation should be careful to incentivize "good" reporting behavior. The question of motivation should also be studied qualitatively, e.g., in the same way that Malinen [2015] studies motivations to contribute to online platforms.

Making sense of unstructured text. Prior work has highlighted the challenge of bridging qualitative and quantitative insights (e.g., DeVos et al. [2022], Deng et al. [2023]). Recent methodological work to this end leverages developments in LLMs (e.g., Rao et al. [2024], Movva et al. [2025], Tamkin et al. [2024], Chausson et al. [2025]), which may prove fruitful. However, we note that even more classical NLP approaches (e.g. as in Robinson [2019], Ayres et al. [2013], Bastani et al. [2019]) have shown to be effective when processing existing (natural-language) report data.

Sequential analysis and online decisionmaking. Statistics has studied sequential testing for decades (e.g., Wald's SPRT [Wald and Wolfowitz, 1948]). More recently, e-values (e.g., Vovk and Wang [2021]) have emerged as a popular framework for sequential approaches; however, as Dai et al. [2025] note, nontrivial extensions for settings more complex than a single hypothesis are open technical problems. Thus, it may be fruitful to explore methods that loosen the stringency of a true sequential hypothesis test, and/or that seek to reconcile the statistical approach with the text-based methods discussed above. In particular, insights from the rich literature in multi-armed bandits and other approaches to online optimization and decisionmaking may offer methodological contributions to handling sequentiality.

6 Discussion

This work seeks to establish *aggregated individual reporting* as a conceptual framework for post-deployment evaluation of AI systems. There are several well-founded criticisms of aggregated

[8] While there are, of course, domain experts from other fields that have insights on the design and implementation of reporting systems, we focus here on computer science-adjacent subfields.

individual reporting; nevertheless, we see these concerns as important considerations for carefully designing and improving AIRs, rather than reasons that they should not be attempted at all.

6.1 Limitations of aggregated individual reporting

We see two major categories of failure modes for AIRs: those arising from crowdsourced reporting as a data source, and those arising from organizational challenges.

(1) Reporting is a fundamentally-flawed source of data. By design, feedback collected by AIRs is "one-sided," in the sense that, and moreover, relies on usage or adoption at scale.

(1a) Individual reports would be too noisy or too biased (e.g., by reports attempting to "game" the system) as to be unusable. This is a serious concern, especially given the existence of (online) crowd behaviors like review bombing [Payne, 2024], and the varying quality of complaints about content moderation algorithms. Anecdotally, for instance, user feedback about moderation is often about explicit content.

(1b) People may not submit reports even if the mechanism technically exists—e.g., if reporting is too burdensome, or if affected populations are unaware of the option to report. Encouraging sufficient participation is also a common challenge across applications that are rely on eliciting data from the public (e.g., study recruitment and retention [Koo et al., 2005]).

Handling both of these concerns is partially an empirical question (in what ways would a reporting system for evaluating AI induce different behaviors? What kinds of report affordances affect incentives?) that can only be understood by analyzing a real-world implementation. On the other hand, as mentioned in 5.2, there are also various research communities that can and should contribute towards these problems.

(2) Organizational challenges may complicate the pathway to downstream action or accountability. One key premise of our proposal is that reports can be empowering because they can effect action, rather than languishing in an online database. However, for this to happen, there are organizational and institutional problems that must be addressed beyond technical and methodological challenges.

(2a) Model developers may not be incentive-aligned as 1st-party operators. Even if problems with the evaluated system are identified, the system developer is not necessarily bound to address them. Though Mozilla's Youtube Regrets study was discussed earlier as a positive example of aggregation [McCrosky and Geurkink, 2021], Youtube never explicitly acknowledged the study as influencing specific choices about their recommendation algorithm. On the other hand, a first-party administrator could evade accountability by avoiding disclosure of findings from the AIR system.

(2b) Running an AIR mechanism as a 3rd-party may be unsustainable. Many third-party audits are foundation-funded (e.g., RegretsReporter by Mozilla), and the path to long-term financial viability is uncertain, which means that many 3rd-party auditors simply cease to exist. For example, the previously-lauded UberCheats browser extension [Marshall, 2021] no longer exists; the system that became Fairfare [Calacci et al., 2025] was originally known as the DriversSeat app [Driver's Seat Cooperative, 2024], but it is unclear whether prior data from DriversSeat was ever used.

These are critical concerns about ideal patterns of implementation, especially when considering which institutions ought to play what roles. We cannot guarantee *a priori* that these failure modes can be avoided, but we hope that by emphasizing the role of organizational factors in 5.1, we can encourage intentional decisionmaking in this regard.

6.2 Calls to action

Despite these limitations, we believe that aggregated individual reporting, as we have defined in this work, is a natural component of the post-deployment evaluation ecosystem. Recent events—and ongoing research—have illustrated that individuals have unique contributions to understanding the contours of AI system behavior. The constellations of individual experience are an invaluable resource not just for model development, but for understanding potentially-unintended societal consequences of already-deployed systems.

Our main call to action, therefore, is for AIRs to be built—and the data collected by them to be analyzed. *Academic researchers* should develop the methodological innovations and empirical analyses necessary for effective implementation; this is an interdisciplinary challenge that spans several subfields of computer science (and beyond), including not just AI/ML but also statistics, human-computer interaction, and law and policy. *Activists and industry practitioners* should begin exploring development of these systems. *Policymakers* should work in collaboration with the aforementioned stakeholders to understand what kinds of legal or policy leverage may be useful.

To translate the AIR approach from a hypothetical strategy to reality, it is essential for the area to mature; addressing the wide range of design and methodological issues outlined in this position is just the first step in coalescing a community to make progress towards this goal. But, at the same time, it is impossible to wait for all the hypothetical kinks to be smoothed: the success of AIRs is a fundamentally empirical question. We believe that it is worth trying to find out.

Acknowledgments

JD is supported in part by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 2146752. BR is generously supported in part by NSF CIF award 2326498, NSF IIS Award 2331881, and ONR Award N00014-24-1-2531. IDR is supported by the Mozilla Foundation and MacArthur Foundation. IYC is supported by a Google Research Scholar Award and Apple Machine Learning Research Grant. The authors thank Nika Haghtalab and Paula Gradu for helpful conceptual discussions and support throughout the process.

References

G. Agostini, E. Pierson, and N. Garg. A Bayesian Spatial Model to Correct Under-Reporting in Urban Crowdsourcing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21888–21896, 2024.

L. Ahmad, S. Agarwal, M. Lampe, and P. Mishkin. Openai's approach to external red teaming for ai models and systems. *arXiv preprint arXiv:2503.16431*, 2025.

U. G. Assembly. Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development, 2024. URL https://documents.un.org/doc/undoc/ltd/n24/065/92/pdf/n2406592.pdf.

I. Ayres, J. Lingwall, and S. Steinway. Skeletons in the database: An early analysis of the cfpb's consumer complaints. *Fordham J. Corp. & Fin. L.*, 19:343, 2013.

J. Baik and H. Sridharan. Civil rights audits as counterpublic strategy: articulating the responsibility and failure to care for marginalized communities in platform governance. *Information, Communication & Society*, 27(5):836–855, 2024.

K. Bastani, H. Namavari, and J. Shaffer. Latent dirichlet allocation (lda) for topic modeling of the cfpb consumer complaints. *Expert Systems with Applications*, 127:256–271, 2019.

J. M. Beaubien and D. P. Baker. A review of selected aviation human factors taxonomies, accident/incident reporting systems and data collection tools. *International Journal of Applied Aviation Studies*, 2(2):11–36, 2002.

C. Bertram. Jean Jacques Rousseau. In E. N. Zalta and U. Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2023 edition, 2023.

S. Bharath. Post on x. https://x.com/Siddharth87/status/1916999455146185022, April 2025.

J. R. Biden. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. 2023.

A. Birhane, W. Isaac, V. Prabhakaran, M. Diaz, M. C. Elish, I. Gabriel, and S. Mohamed. Power to the people? Opportunities and challenges for participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–8, 2022.

D. Calacci, V. N. Rao, S. Dalal, C. Di, K.-W. Pua, A. Schwartz, D. Spitzberg, and A. Monroy-Hernández. Fairfare: A tool for crowdsourcing rideshare data to empower labor organizers. *arXiv preprint arXiv:2502.11273*, 2025.

S. Chausson, M. Fourcade, D. J. Harding, B. Ross, and G. Renard. The insight-inference loop: Efficient text classification via natural language inference and threshold-tuning. *Sociological Methods & Research*, page 00491241251326819, 2025.

W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. Jordan, J. E. Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

Consumer Financial Protection Bureau. Consumer complaint database, June 2012. URL https://www.consumerfinance.gov/data-research/consumer-complaints/.

J. Dai and E. Fleisig. Mapping social choice theory to RLHF. *arXiv preprint arXiv:2404.13038*, 2024.

J. Dai, P. Gradu, I. D. Raji, and B. Recht. From individual experience to collective evidence: A reporting-based framework for identifying systemic harms. *International Conference on Machine Learning*, 2025.

F. Delgado, S. Yang, M. Madaio, and Q. 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*, pages 1–23, 2023.

W. H. Deng, B. Guo, A. Devrio, H. Shen, M. Eslami, and K. Holstein. Understanding practices, challenges, and opportunities for user-engaged algorithm auditing in industry practice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2023.

A. DeVos, A. Dhabalia, H. Shen, K. Holstein, and M. Eslami. Toward user-driven algorithm auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–19, 2022.

Driver's Seat Cooperative. Driver's seat cooperative. https://www.driversseat.co/, 2024.

W. N. Espeland and B. I. Vannebo. Accountability, quantification, and law. *Annu. Rev. Law Soc. Sci.*, 3(1):21–43, 2007.

D. Estlund. Beyond fairness and deliberation: The epistemic dimension of democratic authority. *Deliberative democracy: Essays on reason and politics*, 173:204, 1997.

European Parliament. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act), 2024. Chapter IX, Articles 85 and 87.

M. Feffer, M. Skirpan, Z. Lipton, and H. Heidari. From preference elicitation to participatory ml: A critical survey & guidelines for future research. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 38–48, 2023.

fortheloveoftheworld. This new update is unacceptable and absolutely terrifying. https://www.reddit.com/r/OpenAI/comments/1kasjmr/this_new_update_is_unacceptable_and_absolutely/, April 2025.

N. Fraser. Rethinking the public sphere: A contribution to the critique of actually existing democracy'. *Social Text*, (25/26):56–80, 1990.

Frye. Post on x. https://x.com/___frye/status/1916346474893656572, April 2025.

M. E. Gilman. Democratizing AI: Principles for meaningful public participation. *Data & Society*, 2023.

I. Globus-Harris, M. Kearns, and A. Roth. An algorithmic framework for bias bounties. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1106–1124, 2022.

I. Globus-Harris, D. Harrison, M. Kearns, P. Perona, and A. Roth. Diversified ensembling: An experiment in crowdsourced machine learning. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 529–545, 2024.

V. Gourevitch and J.-J. Rousseau. *Rousseau: the Social Contract and other later political writings*. Cambridge University Press, 2018.

T. H. Davenport. How strategists use "big data" to support internal business decisions, discovery and production. *Strategy & leadership*, 42(4):45–50, 2014.

C. Haendler and R. Heimer. The financial restitution gap in consumer finance: Insights from complaints filed with the cfpb. *Available at SSRN*, 3766485, 2021.

K. Hill. They asked chatgpt questions. the answers sent them spiraling. *The New York Times*, 2025. URL https://www.nytimes.com/2025/06/13/technology/chatgpt-ai-chatbots-conspiracies.html.

E. Ho. Post on x. https://x.com/Eddie_Ho2025/status/1913596114156028013, April 2025.

S. Huang, D. Siddarth, L. Lovitt, T. I. Liao, E. Durmus, A. Tamkin, and D. Ganguli. Collective Constitutional AI: Aligning a Language Model with Public Input. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1395–1417, 2024.

A. Jacobsson, Å. Ek, and R. Akselsson. Learning from incidents–a method for assessing the effectiveness of the learning cycle. *Journal of Loss Prevention in the Process Industries*, 25(3): 561–570, 2012.

P. Kalluri. Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*, 583 (7815):169–169, 2020.

J. Kenway, C. François, S. Costanza-Chock, I. D. Raji, and J. Buolamwini. Bug bounties for algorithmic harms? lessons from cybersecurity vulnerability disclosure for algorithmic harms discovery, disclosure, and redress. *Algorithmic Justice League*, 2022.

R. Klar. Youtube algorithm keeps recommending 'regrettable' videos, study finds. *The Hill*, July 2021. URL https://thehill.com/policy/technology/561788-youtube-algorithm-keeps-recommending-regrettable-videos/.

M. Klee. Ai spiritual delusions are destroying human relationships. *Rolling Stone*, May 2025. URL https://www.rollingstone.com/culture/culture-features/ai-spiritual-delusions-destroying-human-relationships-1235330175/.

M. Koo, H. Skinner, et al. Challenges of internet recruitment: a case study with disappointing results. *Journal of medical Internet research*, 7(1):e126, 2005.

M. S. Lam, M. L. Gordon, D. Metaxa, J. T. Hancock, J. A. Landay, and M. S. Bernstein. End-user audits: A system empowering communities to lead large-scale investigations of harmful algorithmic behavior. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–34, 2022.

K. F. Larson, E. Ammirati, E. D. Adler, L. T. Cooper Jr, K. N. Hong, G. Saponara, D. Couri, A. Cereda, A. Procopio, C. Cavalotti, et al. Myocarditis after bnt162b2 and mrna-1273 vaccination. *Circulation*, 144(6):506–508, 2021.

Laura. Post on x. https://x.com/IndieLauraSDG/status/1913102627837206587, April 2025.

R. Lawler. Mozilla's regretsreporter data shows youtube keeps recommending harmful videos. *The Verge*, July 2021. URL https://www.theverge.com/2021/7/7/22567640/youtube-algorithm-suggestions-radicalization-mozilla.

A. Littwin. Why process complaints: Then and now consumer. *Temple Law Review*, 87:895, 2015.

Z. Liu and N. Garg. Equity in resident crowdsourcing: Measuring under-reporting without ground truth data. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 1016–1017, 2022.

Z. Liu, U. Bhandaram, and N. Garg. Quantifying spatial under-reporting disparities in resident crowdsourcing. *Nature Computational Science*, 4(1):57–65, 2024.

J. Locke. *Two Treatises of Government*. Cambridge University Press, 1988. Originally published in 1689.

S. Longpre, K. Klyman, R. E. Appel, S. Kapoor, R. Bommasani, M. Sahar, S. McGregor, A. Ghosh, B. Blili-Hamelin, N. Butters, et al. In-house evaluation is not enough: Towards robust third-party flaw disclosure for general-purpose ai. *arXiv preprint arXiv:2503.16861*, 2025.

C. Macrae. The problem with incident reporting. *BMJ quality & safety*, 25(2):71–75, 2016.

S. Malinen. Understanding user participation in online communities: A systematic literature review of empirical studies. *Computers in human behavior*, 46:228–238, 2015.

A. Marshall. Gig workers gather their own data to check the algorithm's math. *WIRED*, February 2021. URL https://www.wired.com/story/gig-workers-gather-data-check-algorithm-math/.

M. Marshall, I. D. Ferguson, P. Lewis, P. Jaggi, C. Gagliardo, J. S. Collins, R. Shaughnessy, R. Caron, C. Fuss, K. J. E. Corbin, et al. Symptomatic acute myocarditis in 7 adolescents after pfizer-biontech covid-19 vaccination. *Pediatrics*, 148(3), 2021.

J. McCrosky and B. Geurkink. Youtube regrets: A crowdsourced investigation into youtube's recommendation algorithm. Technical report, Mozilla Foundation, July 2021. URL https://assets.mofoprod.net/network/documents/Mozilla_YouTube_Regrets_Report.pdf.

S. McGregor. Preventing repeated real world ai failures by cataloging incidents: The ai incident database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15458–15463, 2021.

S. A. Mouch, A. Roguin, E. Hellou, A. Ishai, U. Shoshan, L. Mahamid, M. Zoabi, M. Aisman, N. Goldschmid, and N. B. Yanay. Myocarditis following COVID-19 mRNA vaccination. *vaccine*, 39 (29):3790–3793, 2021.

R. Movva, K. Peng, N. Garg, J. Kleinberg, and E. Pierson. Sparse autoencoders for hypothesis generation. *arXiv preprint arXiv:2502.04382*, 2025.

OECD. Oecd ai incidents and hazards monitor (aim), 2023. URL https://oecd.ai/en/incidents.

V. Ojewale, R. Steed, B. Vecchione, A. Birhane, and I. D. Raji. Towards ai accountability infrastructure: Gaps and opportunities in ai audit tooling. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–29, 2025.

OpenAI. Expanding on what we missed with sycophancy. OpenAI Blog, May 2025a. URL https://openai.com/index/expanding-on-sycophancy/.

OpenAI. Sycophancy in gpt-4o: what happened and what we're doing about it, April 2025b. URL https://openai.com/index/sycophancy-in-gpt-4o/.

M. E. Oster, D. K. Shay, J. R. Su, J. Gee, C. B. Creech, K. R. Broder, K. Edwards, J. H. Soslow, J. M. Dendy, E. Schlaudecker, et al. Myocarditis cases reported after mrna-based covid-19 vaccination in the us from december 2020 to august 2021. *Jama*, 327(4):331–340, 2022.

A. Ovadya, L. Thorburn, K. Redman, F. Devine, S. Milli, M. Revel, A. Konya, and A. Kasirzadeh. Toward democracy levels for ai. *arXiv preprint arXiv:2411.09222*, 2024.

W. B. Payne. Review bombing the platformed city: Contested political speech in online local reviews. *Big Data & Society*, 11(3):20539517241275879, 2024.

I. D. Raji, P. Xu, C. Honigsberg, and D. Ho. Outsider oversight: Designing a third party audit ecosystem for ai governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 557–571, 2022.

V. N. Rao, E. Agarwal, S. Dalal, D. Calacci, and A. Monroy-Hernández. Quallm: An llm-based framework to extract quantitative insights from online forums. *arXiv preprint arXiv:2405.05345*, 2024.

B. Recht. A bureaucratic theory of statistics. *arXiv preprint arXiv:2501.03457*, 2025.

P. Reviews. Post on x. https://x.com/ReviewsPossum/status/1917292836082397355, April 2025.

S. Robinson. Temporal topic modeling applied to aviation safety reports: A subject matter expert review. *Safety science*, 116:275–286, 2019.

J. Scott. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. The Institution for Social and Policy Studies. Yale University Press, 1998. ISBN 9780300078152. URL https://books.google.com/books?id=PqcPCgsr2u0C.

H. Shen, A. DeVos, M. Eslami, and K. Holstein. Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–29, 2021.

T. T. Shimabukuro, M. Nguyen, D. Martin, and F. DeStefano. Safety monitoring in the vaccine adverse event reporting system (VAERS). *Vaccine*, 33(36):4398–4405, 2015.

J. A. Singleton, J. C. Lloyd, G. T. Mootrey, M. E. Salive, R. T. Chen, S. Ellenberg, S. Rastogi, C. Krueger, M. Braun, R. Wise, et al. An overview of the vaccine adverse event reporting system (vaers) as a surveillance system. *Vaccine*, 17(22):2908–2917, 1999.

P. Slattery, A. K. Saeri, E. A. Grundy, J. Graham, M. Noetel, R. Uuk, J. Dao, S. Pour, S. Casper, and N. Thompson. The ai risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. *arXiv preprint arXiv:2408.12622*, 2024.

T. Sorensen, L. Jiang, J. D. Hwang, S. Levine, V. Pyatkin, P. West, N. Dziri, X. Lu, K. Rao, C. Bhagavatula, et al. Value kaleidoscope: Engaging AI with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19937–19947, 2024a.

T. Sorensen, J. Moore, J. Fisher, M. Gordon, N. Mireshghallah, C. M. Rytting, A. Ye, L. Jiang, X. Lu, N. Dziri, et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024b.

A. Tamkin, M. McCain, K. Handa, E. Durmus, L. Lovitt, A. Rathi, S. Huang, A. Mountfield, J. Hong, S. Ritchie, et al. Clio: Privacy-preserving insights into real-world ai use. *arXiv preprint arXiv:2412.13678*, 2024.

The Next Web. Youtube recommends videos that violate the platform's own policies, study finds. *The Next Web*, July 2021. URL https://thenextweb.com/news/youtube-algorithm-recommends-videos-that-violate-platforms-policies-mozilla-study.

V. Vovk and R. Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

A. Wald and J. Wolfowitz. Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, pages 326–339, 1948.

Williawa. Post on x. https://x.com/williawa/status/1916935712550551743, April 2025.

G. Witberg, N. Barda, S. Hoss, I. Richter, M. Wiessman, Y. Aviv, T. Grinberg, O. Auster, N. Dagan, R. D. Balicer, et al. Myocarditis after Covid-19 vaccination in a large health care organization. *New England Journal of Medicine*, 385(23):2132–2139, 2021.

A. W. Wu, P. Pronovost, and L. Morlock. Icu incident reporting systems. *Journal of critical care*, 17 (2):86–94, 2002.