PNAS PERSPECTIVE OPEN ACCESS

How public involvement can improve the science of AI

J. Nathan Matias[a,b,1] and Megan Price[c]

Edited by Moshe Vardi, Rice University, Houston, TX; received April 18, 2025; accepted September 22, 2025

As AI systems from decision-making algorithms to generative AI are deployed more widely, computer scientists and social scientists alike are being called on to provide trustworthy quantitative evaluations of AI safety and reliability. These calls have included demands from affected parties to be given a seat at the table of AI evaluation. What, if anything, can public involvement add to the science of AI? In this perspective, we summarize the sociotechnical challenge of evaluating AI systems, which often adapt to multiple layers of social context that shape their outcomes. We then offer guidance for improving the science of AI by engaging lived-experience experts in the design, data collection, and interpretation of scientific evaluations. This article reviews common models of public engagement in AI research alongside common concerns about participatory methods, including questions about generalizable knowledge, subjectivity, reliability, and practical logistics. To address these questions, we summarize the literature on participatory science, discuss case studies from AI in healthcare, and share our own experience evaluating AI in areas from policing systems to social media algorithms. Overall, we describe five parts of any quantitative evaluation where public participation can improve the science of AI: equipoise, explanation, measurement, inference, and interpretation. We conclude with reflections on the role that participatory science can play in trustworthy AI by supporting trustworthy science.

AI | participatory research | policy | evaluation | citizen science

In the spring of 2024, nursing professionals represented by the California Nursing Association protested outside Kaiser San Francisco Medical Center to draw attention to what they describe as "unproven AI that could put patients at risk." The nurses worried that AI technology was being implemented in a rush, with insufficient transparency regarding its uses and how it would help patients or staff. They also argued that a new system used to determine levels of patient care did not adequately account for the time-consuming and sensitive work that nurses do. At the protest, union leaders did not demand that the system be abolished; instead they demanded more input into the design and evaluation of AI systems (1). If the union achieves their demands, both sides will need a way to incorporate practitioner expertise into the science of AI evaluation.

When nurses at the San Francisco hospital expressed concerns about AI, they were referring to a class of software systems that were broadly probabilistic in nature (2). Such systems have so many different underlying technical mechanisms (from deep learning to markov models to spreadsheets enhanced with generative AI) that scholars and policymakers alike have struggled to classify the underlying technology (3). In practice, Kaiser Permanente, like other organizations, has long invested in digital systems for managing information, coordinating staff, guiding decisions, and computing complex logistics (4). Where prior, deterministic computer systems have been encoded with rules to predictably carry out organizational agreements about how work should be structured, newer "AI" systems are trained on historical data and simulations to make inferences that may not be explainable with institutional rules (5). In clinical settings, for example, AI has been proposed for diagnosis (6), scheduling (7, 8), prediction of patient deterioration in hospitals (9), and messaging with patients (10). Similar transitions are underway in nearly every organizational setting, including education (11), law enforcement (12), judicial sentencing (13), insurance (14), and scientific peer review (15) to name a few.

One demand from the protesting nurses included a request for a seat at the table of AI evaluation. This demand was fundamentally about measurement, statistics, and theories of science: Can we generate more reliable scientific evaluations of AI systems and their use in context when those evaluations include perspectives from lived-experience experts such as nursing professionals or patients? Or is engagement with affected stakeholders a necessary concession that reduces the objective quality of AI evaluations in exchange for wider acceptance? Similar debates about rigor have occurred in other areas where people have adopted participatory methods to advance science and evaluate technologies (16, 17).

External demands for public involvement in AI evaluation run parallel to calls from scientists to address shortcomings in this same science. Because AI systems are trained on human behavior and interact with complex human-systems, scientists have struggled to reliably predict the total outcomes of an AI system in use (18). To address this weakness, scientists have turned to publicly engaged methods to broaden the capacity of science to observe and test AI systems in their specific use cases (19–22).

Questions about the role of participatory science in AI evaluation have animated the work of both authors for

Author affiliations: [a]Department of Communication, Citizens and Technology Laand staff buy-b, Cornell University, Ithaca, NY 14853; [b]Coalition for Independent Technology Research, San Francisco, CA 94188; and [c]Human Rights Data Analysis, Los Angeles Group, CA 90012

Author contributions: J.N.M. and M.P. designed research; performed research; and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: nathan.matias@cornell.edu.

Published November 14, 2025.

over fifteen years across a wide range of systems and situations: generative AI product evaluation, human rights investigations, audits of decision-making algorithms, analyses of predictive policing systems, and field experiments in collective human–algorithm behavior. This article shares that experience and scholarship by arguing how public involvement can improve the science of AI. We open by summarizing the challenge of evaluating mission-critical AI systems as a sociotechnical challenge. We then summarize leading models for public participation in scientific research, failures that arise from disincluding the public, and concerns about expertise that sometimes lead to those failures. Finally, using examples from a range of fields including our own work, we summarize five elements of AI evaluations that can be made more scientifically rigorous by engaging with the public.

Technical, Social, and Political Challenges in Mission-Critical AI Systems

Many of the challenges of deploying trustworthy AI in high stakes contexts such as hospitals constitute what scholars describe as sociotechnical problems: technology questions with deeply embedded social and political aspects (23, 24). Consider the case of clinical diagnoses. On the surface, such tasks might seem like straightforward acts of inference based on a set of inputs. Machine learning models have demonstrated remarkable accuracy in diagnosis based on medical imaging, where developers have access to large datasets of consistently measured and labeled observations (25–27). Yet many diagnoses depend on social interactions between humans under time constraints, such as self-reports of pain that can be error-prone (28). Beyond any specific inference, the concept of a diagnosis is also a bureaucratic and institutional endeavor that is deeply embedded in the history, economics, organizational practices, and regulation of medicine, shaped by local work practices and international treaties at all scales (29). At the clinical level, organizations have developed workflows to determine who is responsible to make a diagnosis under what conditions (6), legal systems have agreed on who can be held responsible for misdiagnoses (30), and public health institutions have standardized how diagnosis and cause of death data should be aggregated and shared between countries (29, 31). Consequently, errors and accuracy in medical systems, like any complex organizational system, are sociotechnical in nature, deeply integrating elements of statistical inference with interpersonal, organizational, and political concerns.

Many of these fundamental insights about the social layers of AI systems in context have been developed by qualitative and critical scholars. In this paper, we focus on the implications of this work for forms of research in computing, statistics, and the social science that are largely quantitative.

Reliability. Scientists and advocates have developed multiple subfields to document and address the reliability and risks of AI systems. In the most simple formulation, researchers optimize a model for one or more outcomes relevant to their goals, seeking to resolve mathematically describable problems of accuracy, precision, and error, especially for out of sample inference (32). For example, researchers studying fairness have made substantial progress on paradigms for observing bias in judgments, identifying the reasons for that bias, and addressing it in the underlying model (32, 33).

Yet as statisticians, computer scientists, and social scientists have observed, the total outcome of a model's reliability is embedded in multiple layers of social and psychological factors that surround the decisions an AI model supports. For example, extensive scholarship has documented these social layers in the case of pretrial risk assessment in criminal justice (34). At the mathematical layer, organizations choosing an AI system must choose among many possible mathematical definitions of accuracy and fairness (32). Further social factors shape the layer involving the data that a model is trained on. When organizations turn to AI in the search for improvements in how the organization operates, they often face the problem that their administrative records include systematic errors and represent different values that they are hoping to transform through AI (32, 34). Once a model is trained and adopted, an interpretation layer filters the outputs of a model through social and psychological factors among judges, juries (35, 36) and the public as they make sense of those outputs (37). Beyond individual psychology, AI outcomes are also subject to an organizational layer, since legitimate decision-making processes need to follow the regulatory requirements of the court system. As statisticians have pointed out, technical decisions about outcome variables and standards of reliability depend on all of these social layers (34). In a final layer of temporal validity, the reliability of AI models tends to degrade over time as conditions change, diverging from the social and institutional situations that produced the model's training data in the first place (38).

Performance in Context. Even where evidence on AI reliability is available, efforts to introduce AI into organizations wrestle with the fundamental problem of mismatches between system design and "street-level" implementation (39). Automated systems can in theory build trust and minimize errors by reducing the discretion of decision-makers (40). In practice, those advantages vary with the setting where a system is deployed. In highly context-dependent domains such as public health, criminal justice, and child welfare, for example, reductions in discretion can override important expertise and context that organizations rely on to make systems workable and humane (41).

The expected efficiencies promised by AI creators have also been unrealized in settings where a new technology was not adequately incorporated into existing workflows. In a review of studies examining whether technologies saved or cost health care workers time, researchers concluded that some of the main causes of costs to staff time were related to workflow, usability, and staff buy-in (42).

Security Risks. Finally, extensive cybersecurity research has found that high-stakes AI systems include multiple layers of security risks from general flaws, security attacks, and privacy attacks (43). Security risks are especially acute for critical infrastructure with significant implications for life and livelihoods. In healthcare for example, breaches of health records create high risks for patients, and healthcare delivery organizations face strong pressures to comply with ransom demands. Even before the release of ChatGPT in 2022, health data breaches and ransomware attacks had collectively grown to hundreds of incidents per year in the

United States, leading to data leaks affecting tens of millions of patients, electronic system downtime, cancellation of scheduled care, and diversion of ambulances (44, 45). The introduction of AI creates further surface areas for security risks that scientists are still seeking to understand. To name a few, general flaws include errors from confabulation (sometimes called hallucination), bias, and forgetfulness (43). Security attacks force a model to alter its behavior through data poisoning, prompt injection, or jailbreaking attacks. Privacy vulnerabilities occur when a system inadvertently leaks private information about the people and institutions in their training datasets (43). The more general-purpose a system is designed to be, the harder it is for designers to anticipate and prevent these attacks (5).

Computer scientists have developed participatory methods for identifying such attacks, including bias bounties and red teaming, where people imagine possible attacks, attempt them, and report the results. Analyses of these still in-development methods have argued that they can help identify some vulnerabilities but are not yet able to provide reliable evidence about the overall security of a system or compare the security of different AI systems (19, 46, 47).

Transparency. Despite much progress on the science of evaluation, trustworthy evidence about even the most basic statistical reliability of AI systems is rarely available to the customers, users, or parties affected (5). The problem of transparency is not unique to AI. Healthcare and many other applied fields face a general challenge, where ignorance about the reliability or outcomes of their systems is in the self-interest of many parties (48, 49). Indeed, the general notion of scientific falsification rests on philosophical arguments concerned with producing knowledge that the public can use to reject the harmful or unreliable exercise of power (50).

In practice, interested parties tend to direct the production, publication, and interpretation of evidence to serve their interests. Within medicine for example, scientists have observed that without regulatory obligations to publish all trials, drug makers tend to publish favorable clinical trials and suppress clinical evidence that could generate liability or reduce profits (51). This pattern has also been observed in the field of algorithm auditing, where a file drawer problem in bias audit results is consistent with strategic violation of transparency laws to avoid discrimination lawsuits (3, 52). The strategic deployment of science for regulatory evasion is to be expected. Scholars who study consumer protection and environmental governance have described transparency as a coevolutionary race, where multiple actors compete to accomplish their goals by advancing science, preventing it, and evading ever-novel barriers to transparency. In this model, successful governance arises from an equilibrium in these knowledge-generating efforts (53, 54).

Models of Public Involvement in AI Research

How can the science of evaluation incorporate the social and organizational layers of system performance that computer scientists are not always trained to account for? Across the research endeavor, participatory research is an approach that includes a consequential role for people who bring capabilities and expertise beyond that of professional scientists (55). Participatory methods, which have been developed for nearly every imaginable mode of inquiry (56), have become essential to science and evaluation in many fields including biology, environmental science, toxicology, astronomy, and many others (55). These scientific developments have run in parallel to efforts in participatory governance in urban planning (57), industrial labor (58), design (59), healthcare (60), and politics (61, 62). This approach has been called civic science (63, 64), citizen science (55), engaged research (65, 66), action research (58), user-controlled research (67), and participatory science (68), to name a few. In this section, we describe leading models of public involvement in science and governance, alongside how they are being used in the science of AI.

Contributory Science. Many AI researchers have adopted a model of contributory science (55), where members of the public contribute observations and judgments that are used to train and evaluate AI systems (69). Contributory projects are especially useful for scientific problems with high variance, where members of the public can cross the earth to make observations, report context-specific variations in model performance, or contribute their own experience of variation in human society. In this common model, scientists define the goals, outputs, and processes of research and create simple pathways for the public to contribute.

Whether scientists pay participants or ask them to volunteer, contributory projects tend to minimize the agency and discretion of participants in pursuit of valid, commensurate observations that can be aggregated for analysis. In AI research, contributions are sometimes taken without the public's consent, awareness, or opportunity to refuse, leading to high profile class action lawsuits, the withdrawal of datasets, and sometimes even the destruction of AI models (70, 71). Scholars and advocates have also questioned the treatment of contributors, since paid workers are sometimes compensated poorly for jobs that reportedly entail serious mental health risks (72, 73). Yet in areas such as real-time forecasting of bird migration (74), scientists have trained AI models on data donations from over a million contributors who find purpose, community, and recreation from contributing data (75).

As machine learning models are trained to replicate tasks that scientists once relied on the public to carry out, researchers have raised concerns about potential collapses in AI reliability and public contributions. When AI-generated material is incorporated into training data, AI model reliability can degrade substantially (76). In parallel, researchers fear collapses in human participation if contributors feel that they are merely training or competing with automated systems (77). In both cases, successful contributory projects that integrate AI depend on humans offering observations that are unique to their context, situation, and person.

Cocreated Science. Some participatory AI research takes a cocreated or user-controlled form, where affected communities collaborate with or direct scientists to carry out research (55, 59, 67, 78). In this form, affected communities often exercise leadership in the selection of problems, questions, and hypotheses (55). To support cocreation, scientists in fields ranging from epidemiology and microbiology to economics have created and evaluated processes for producing rigorous science in a cycle of collaboration across problem identification, inquiry, and interpretation (79–81).

Within AI research, cocreated science tends to focus on design, evaluation, or both. Many computer scientists and civil society organizations have designed AI systems together with affected communities (20, 82–85). In this process, AI designers host deliberations with affected parties to collect requirements for inference and deployment, choose the data that a system will train on, and contribute to the training of the system. Such cocreation processes have been developed for end-user systems and foundation models alike (86, 87). Where different groups have competing interests in zero-sum situations (organ donation for example), these consultations and decision-making processes help stakeholders negotiate and agree on the trade-offs (88).

Other cocreated research focuses on independent evaluation of "black box" AI systems where communities or purchasers have been excluded by vendors from system design. Such studies, whether called participatory evaluation, audit studies (21, 22, 89, 90), impact assessment (91, 92), or red teaming (93, 94), have been taken up by journalists, civil society, and academic groups to study the accuracy, bias, safety, and security of AI systems. In cocreated studies, scientists work alongside affected parties to develop details of measurement, study design, and interpretation (95). While cocreated evaluation is sometimes welcomed by the makers of AI systems, the relationship can also be adversarial (18).

Participatory Governance. Beyond the scientific work addressed in this article, some participatory AI research engages the public in the governance and oversight of AI products (82). In this model, affected communities contribute to the creation and implementation of organizational or government policies concerning AI. Research on participatory AI governance largely follows the contours of wider work on participatory governance in political science and statistics. Some projects survey public opinion about AI (96, 97). Others convene statistically representative citizens assemblies to advise technology vendors, purchasers, and government regulators (98–101). These efforts at soliciting input from the public are sometimes combined with efforts to use AI systems to support public participation in democratic deliberation (84, 102). Other projects study the actions of labor unions and social movements who develop their own pathways to shape AI governance (103, 104).

Lived-Experience Expertise. Across all of these models, researchers seeking to characterize the scientific validity of participatory research need to conceptualize the value that participants contribute to an inquiry. In public health and social work, participatory researchers use variations on the term "lived-experience expertise" to refer to the knowledge and authority that people bring to scientific inquiry beyond formal training or credentials (105, 106). This expertise includes both that of workers and patients. The unique knowledge and standpoint of such experts may derive from aspects of a person that they cannot choose or change, as is the case in participatory audits of algorithm discrimination (22, 107). In health and medical research for example, this expertise can be derived from the experience of certain medical conditions (108), even if scientists have not yet defined or classified a condition (109). In environmental research and information security, people possess situated expertise associated with their location in the physical world (110), their availability at specific moments of time (111), and network position in digital networks (112). Scholars have observed that as people participate in multiple projects over time, their expertise in professional science also grows (113), and some participatory projects offer recruitment pathways for professional and scientific training (105).

Concerns About Participatory Science

In recent years technologists and policy experts have acknowledged the inseparability of technical and social questions in AI, calling for scientists to acknowledge these considerations and integrate them into the design and deployment of systems (86, 88). One reason this integration has proven difficult to accomplish is that scientists reasonably question whether participatory research can produce high quality science (114). When surveyed and interviewed about their views on participatory research, many scientists report concerns that public participation could reduce the scientific value of their work (115). In this section, we summarize four common concerns and the state of research on those concerns.

Concern: Divides Between Theory and Application. Historians of science have observed that the scientific search for generalizable knowledge is sometimes accompanied by a dichotomy of basic and applied research (116). In AI, this desire for general knowledge runs parallel to financial incentives to create products that scale to the widest number of contexts in what has been called a "ladder of generality" (5). In parallel, computer science is structured with a prestige hierarchy where subfields that are more demographically diverse and more concerned with sociotechnical issues are perceived to be lower status (117). Since participatory science is rooted in specific contexts, it can be seen as undesirable, low-status work that is unlikely to advance theory or maximize profits (20).

For policymakers and institutions such as hospital systems, applied research is valuable in its own right, whether or not a study contributes to scientific theory (118). Nonetheless, researchers who wish to advance theory by attending to context can do so by drawing from long traditions in computing, social science, and the history of science. In experimental psychology for example, traditions of pragmatism (119), action research (58, 120), and full-cycle research (121) have advanced the theoretical strength and scientific reliability of their fields through empirical work in context (122). In computer science, the field of participatory design has derived general processes and principles for design that account for variation in context (123). Across the academic endeavor, historians have argued that many fundamental innovations in science and industry have started with a pragmatic question through use-inspired basic research (116).

Concern: Subjectivity Weakens Science. Scientists sometimes question whether interested parties are able to exercise the disinterestedness and objectivity essential to the scientific endeavor (124). They rightly wonder how to arrive at reliable answers when affected parties (such as a nurse's union) have a vested interest in the system. Yet subjectivity is highly compatible with expertise, as scientists acknowledge every time they value the knowledge of commercial researchers with a financial interest in their own products (125). More broadly, science within democracies depends on "multiple and conflicted subjectivities" to negotiate more complete

understandings of complex issues (124). For example, scholars of decision-making algorithms have observed that when institutions are making decisions about consequential systems, the work of building trust involves including and balancing interests rather than omitting them entirely. The design of algorithms for kidney allocation in the United States for example included input from people who could be eligible for kidneys and not just scientists with no stake in the outcome (88).

Concern: Public Participation Reduces Reliability. Scientists sometimes also have a concern that empirical work involving people who are not professional scientists will lack the precision or rigor that science requires. These include concerns about sample diversity and measurement error. Since AI research has long relied on human data, digital labor, and crowd-sourced data under the term human computation (72, 126), these concerns are already the subject of active debates and reforms in AI research.

Problems with sample bias have long plagued AI research, whose reliance on widely available digitized information has created problems of fairness and misrepresentation in AI models (127–129). Furthermore, research studying the potential harms of digital technologies has itself been constrained by strong cultural, linguistic, and geographic bounds (130). Scientists who collaborate with the public typically respond to these concerns in two ways. First, they advance purposive sampling of public participation as a corrective to the sample biases propagated by more conventional science (131). Second, in cases where sample bias is an issue scientists apply statistical techniques to account for that bias in the data (132, 133).

Many participatory science projects reckon with problems of measurement accuracy due to variation in self-reports or the prohibitive cost of more precise equipment. The most common practices to improve measurement reliability involve training, validation by experts, and statistical modeling of systematic error (16). In areas such as medical research where self-reports are unavoidable, scientists have created hybrids of self-report and AI models to improve the reliability of observations (28). Scientists can also work with the public to measure AI directly and precisely with measurements and interventions by humans and software working together (107).

Concern: Cost, Pace, and Scale. Scientists are sometimes reluctant to conduct participatory research due to the costs, pace, and scalability of partnering with the public. These are legitimate concerns, since collaborations require training and resources to support the community-engagement process. Consequently, participatory projects can have longer timelines than projects that do not include the public (134). Furthermore, technologists focused on broad, rapid, adoption at low cost have observed that public engagement increases the costs of scaling a system to many settings (20).

As the nurses on strike demonstrated, this quest for premature scale offers the appearance of speed rather than the reality. Technologists often move fast by ignoring the situations and experiences of people who must then live with the consequences. The resulting errors, costs, delays, and irreversible harms can exceed the benefits of producing products or papers a few months more quickly (135). In this article, we focus on the production of rigorous and reliable science, which takes time, resources, and care to get right.

Five Areas Where the Science of AI Evaluation Benefits from Public Involvement

Public participation is sometimes seen as a necessary compromise with the social and political complexities of introducing a more ideal system. Yet public involvement and oversight of highly technical systems can also improve the quality and rigor of the underlying science. Here, we summarize five specific areas where involving lived-experience experts can improve quantitative evaluations of AI.

These five areas represent key elements in the larger arc of a quantitative scientific evaluation. When a project is being imagined, equipoise is a principle that brings people together to agree on a research question in the first place. Next, study designs often hinge on questions of measurement: defining what is important, what can be measured, and how those measurements relate to what matters to stakeholders. When designing the structure of a study design, computer scientists often seek explanations about an AI system, a form of knowledge that enables reasoning about counterfactuals in system design or use—akin to understanding the underlying mechanisms of a psychological effect. Within the overall structure of a study, evaluations also involve inference: the statistical methods by which we draw conclusions, alongside the factors that we include (or exclude) in the models to support those conclusions. Finally, research concludes with interpretation, the process by which scientists and stakeholders make sense of the statistical results.

In this section, we consider how participatory methods can improve equipoise, measurement, explanations, inference, and interpretation of scientific evaluations of AI.

Equipoise. Participatory science can help negotiate disagreements about AI systems by including affected parties in the design of evaluations. When science is invoked by a group of people who disagree with each other or are uncertain over a course of action, it allows those people to commit to a shared process and accept the outcome (136)–a condition called equipoise (137). In medical science, this principle is used to balance the interests of scientists seeking new discoveries with the interests of patients seeking the best treatments (138).

Applied to AI, the principle of equipoise is invoked whenever stakeholders with differing interests determine the criteria by which a system will be evaluated. In healthcare, the allocation of scarce organs for transplant illustrates the problem of convening competing stakeholders. In this setting, each adjustment to an allocation algorithm could privilege one group over others, whether they are older, younger, have more preexisting conditions, or are geographically further from a hospital. Yet designers of the US Organ Procurement and Transplant Network have developed a process for discussion and evaluation of the allocation algorithm that the relevant parties have been willing to join and accept (88).

Importantly, the success of a process does not answer a question forever. Instead, success is marked by an acceptance to abide by agreed-upon standards and a commitment to an evidence-based process for resolving further disputes. This has notably been the case in deliberations over the allocation of donated organs, where stakeholder groups have carried out ongoing negotiations and adjustments of organ allocation criteria (88).

Beyond medicine, equipoise has played a key role in the history of labor disputes when unions and employers have established bureaus of standards that include scientists from both sides to negotiate the scientific details of union contracts (139). Equipoise has also been invoked in computer science and technology management to describe a commitment to resolving disagreements with evidence rather than simply following the preferences of leaders with the most to gain (140, 141). While no single study can conclusively settle competing interests forever, political scientists have observed that scientific evaluation can help achieve an equilibrium in this "coevolutionary race" by improving the state of knowledge about a complex system over time (54).

Equipoise is also possible in the most contested situations. Through her work leading the Human Rights Data Analysis Group (HRDAG), one author has seen the power of equipoise in the midst of determining accountability following armed conflict. HRDAG has been involved in multiple truth commissions, entities created as a way for adversarial parties to agree to a process of uncovering information, establishing a historical record, and identifying a path forward for postconflict communities. For example, after the peace agreement in Colombia in 2016, the Colombian Truth Commission collaborated with HRDAG specifically to use AI tools to integrate multiple disparate datasets describing over thirty years of conflict to generate a single, credible set of estimations of violence (142). The goal of the collaborative project was to create a unified quantitative set of findings to serve as a common starting point for decisions regarding accountability and amnesty. In Guatemala in the late 1990s, HRDAG encountered equipoise in action in its role with the Guatemalan Truth Commission. In February 1999, the commission published a report that included estimates by HRDAG of people killed or disappeared during the armed conflict between 1962 and 1996 (143). One month later, U.S. President Bill Clinton cited these findings in an apology to the people of Guatemala, saying that "support for military forces or intelligence units which engaged in violent and widespread repression of the kind described in the report was wrong" (144).

Measurement. AI research has long relied on public participation to improve the quality and precision of measurement in training sets (5, 128, 129). Many of these endeavors compute averages of multiple judgments to estimate the ground truth of a labeling task. But for social and organizational constructs such as race, organizational policies (145), or even a medical assessment of pain (28), variation among the general public reflects important differences between value systems, social contents, and attitudes (146).

In one author's work with investigative journalists at the Invisible Institute, they saw first-hand how public health categories that community members assigned to police interactions differed substantively from the official categories assigned by the police themselves. Specifically, individuals in Chicago can choose to submit a formal complaint record following interactions with the police. The department then reviews these complaints and groups them according to the primary category of the complaint, such as illegal search or verbal abuse. The police department then publishes aggregate counts summarizing these categories. But when the Invisible Institute gained access to the underlying complaint records themselves, including narrative descriptions of interactions with police officers, they relied on community members to review a sample of the records, assign the records categories that were meaningful to them, and collaboratively developed a machine learning model to label all the records. The result was the discovery that the police's internal process of choosing one primary category for each complaint often obscured additional allegations contained in the narrative. For example, community reviewers found more than a dozen documents the police categorized as complaint of an illegal search, which also described allegations of sexual violation (147, 148). If statisticians only considered the official police categories, their analyses would be inaccurate. The meticulous review by community members brought to light additional information initially hidden in the existing records.

In healthcare, scientists have improved the reliability and precision of hard-to measure constructs with the help of contributory citizen science initiatives (149). In neuroscience and mental health for example, smartphone applications have broadened sample diversity beyond narrow clinical populations and collected information about how people classify their own experiences in ways administrative records might miss (150). In cases such as pain scales, researchers have been able to use AI to model and reduce systematic biases (28). Researchers have also reduced bias by purposively sampling juries involved in participatory evaluation to match the demographics of stakeholders in a given situation (151). In other cases, variation between and within stakeholder groups can provide an early warning that scientists may need to reconceptualize the constructs we seek to measure (146).

Explanation. Evaluation in practice often involves a process of deliberation and adjustment. To support potential adjustments, science needs to provide enough knowledge to forecast the outcomes of a proposed change to the software, user experience, or organizational processes of a system (88). Consequently, stakeholders seek evaluations that offer explanations—counterfactuals about the possible effects of changes to a system. Explanations can include reports from interpretable AI systems about how they arrived at their outputs (152), patterns in the observed behavior of "black box" systems (32), psychological mechanisms explaining human–algorithm interaction (153), or sociological explanations for how algorithms and institutions interact (154).

Lived-experience experts can make fundamental contributions to scientific explanations of human–algorithm behavior, especially at a time when scientists often lack reliable explanations for these interactions (18). In social psychology, such experts have long contributed to hypothesis development and the design of theory-advancing experiments, based on intuitions derived from observation and experience (122). Computer scientists have called these intuitions "folk theories" for how AI systems react to human inputs (155).

In practice, lived-experience experts can improve the science of AI by shaping the design of studies that generate explanations for emergent behavior. Consider for example, a field experiment that investigated the effect of human fact-checking on the promotion of articles by recommender systems (156). In this study led by one of this article's authors, leaders of a 14 million subscriber news discussion community were seeking a way to manage unreliable news sources. Community leaders developed a theory that

fact-checking interventions could affect more than just the beliefs and behaviors of community members. They hypothesized that since the prominence of news articles was decided by a black box algorithm based on observations of community behavior, that influencing humans could have a second-order effect on the algorithm. To investigate this novel theory, participants made suggestions of variables to measure, possible confounding factors, and intervention details that might be most influential. What started as a community objection to the study had become a central contribution to a sociotechnical theory of human–algorithm behavior (95, 156).

Inference. Scientists also benefit from community input when making decisions about inference, the process for drawing conclusions from evidence collected in a scientific evaluation. In quantitative AI research, this could include a model's performance on a benchmark compared to other systems or the result of a statistical test for fairness. The results of such tests are often presented as evidence that a given model is more accurate, less biased, or less prone to errors than humans or a competing model. Yet many such tests exist, and nuanced differences in the goals or usage of a system lead to very different choices in evaluation metrics (32). Consequently, while the design of inference methods for evaluating AI systems requires many years of training, even the most sophisticated methodologist has much to learn from affected parties, both in the formulation of problems (157) and the choice of analysis techniques.

Researchers at the HRDAG and the American Civil Liberties Union (ACLU) learned the value of stakeholder input for inference when estimating bias in a child welfare algorithm in Pennsylvania. The Allegheny Family Screening Tool (AFST) was used to screen calls about alleged child neglect by predicting the likelihood that the agency will remove a child from the family in the next two years. When developed, the system was evaluated for fairness by the scientists who created it (158). But after families began to raise questions about whether and how the AFST was a factor in decisions to remove children from their care, the US Department of Justice began an investigation into the tool's performance (159).

Shortly after the investigation opened, HRDAG and the ACLU conducted an independent analysis of the AFST. They found a mismatch between the way the tool was used and the way it was tested by its creators. The system's developers tested its accuracy and fairness for judgments about individual children and not families. The designers also chose an evaluation metric (AUC) that treated each prediction as an independent judgment without accounting for how the system would be implemented. In practice, the algorithm was used to rank children predicted to be removed on a priority list above others who were predicted not to be removed (158). Bipartite ranking violates assumptions of the AUC metric used by the original evaluators, potentially leading scientists to conclude that an AI system is fair when it is not. HRDAG worked with an interdisciplinary team with extensive knowledge of the child welfare system to choose inference methods that matched the system's use in context—settling on cross-AUC, a method designed to evaluate bipartite ranking (160). In an analysis that was more closely matched to the actual usage of the algorithm in a priority ranking, they found systematic bias against Black families that the original system evaluation missed (161).

While inference techniques for evaluating the AFST are just one part of a larger debate over predictive screening tools (162), this story illustrates how direct input from people with lived-experience expertise can fundamentally improve methods of inference in the science of AI. Because affected families voiced their concerns about something that seemed wrong to them, scientists revisited the published evaluations of the system, found a faulty method of inference in those evaluations, and were able to choose a more appropriate analysis technique.

Interpretation. Affected communities also contribute unique insights to the interpretation of scientific findings and to the deliberations that science supports. When discussing interpretation and deliberation, we are not focused on whether a given model of participation leads to an optimal deliberative outcome among stakeholders (57). Instead, we focus on ways that public involvement can improve the precision of scientific interpretations for scientists and stakeholders alike. This is especially important for scientists, since unrecorded details from the context or implementation can weaken the reliability, replicability, or external validity of study results (163).

Practices such as "participatory hypothesis testing" (164) and "community debriefing" (95) improve scientific validity by inviting affected communities to review preliminary findings, propose explanations for the results, and suggest unrecorded confounding factors. One of the authors has facilitated these reflective practices in multiple large-scale online conversations with thousands of contributors. In one study, participants in an online forum suggested ideas to include in a statistical model that estimated why hundreds of communities joined a strike against a technology firm (164). In other studies, participants shared their on-the-ground experiences of algorithms during a given study and reflected on how those qualitative observations might nuance the interpretation of findings (95). In one case, participants observed that the algorithm-maker had changed how the system worked after the study was underway, an issue that can undermine study validity (165). In response, the community pinpointed the time of the change, supporting adjustments that preserved the validity of the final analysis (156).

Participatory science also grows science by inspiring others to conduct replications. Because the behaviors of AI systems and humans are context-dependent, a single evaluation in one setting cannot reliably predict a system's performance elsewhere without replications (18, 166). Consequently, someone learning the result of an evaluation in a different context should ideally replicate the study rather than unquestioningly apply the results. For simple systems, public curiosity about AI can lead many people to conduct their own evaluation when they learn about a novel result (22). This pattern of contagious cross-validation (167) has also been observed for more complex methods. On the global encyclopedia Wikipedia for example, so many different language communities have independently analyzed the platform's machine learning systems that the organization developed a data repository for querying evaluations across languages and AI system versions (69). These replications then provide further information to future readers on the context dependence of the original result.

Conclusion

In the spring of 2024 when the California Nursing Association (CNA) protested the use of insufficiently tested AI systems, they did not take a position against the use of all AI tools. Instead, they insisted that their expertise be taken into consideration when evaluating and adopting these tools, with CNA president Michelle Gutierrez stating, "Nurses are all for tech that enhances our skills and the patient care experience… We demand that workers and unions be involved at every step of the development of data-driven technologies and be empowered to decide whether and how AI is deployed in the workplace" (168). That same spring, a group of researchers at Kaiser Permanente set out to evaluate a generative AI system for clinical notes across all of the company's geographic regions. Their peer-reviewed case study does not report full evaluation results or indicate whether union representatives were included in the evaluation. Yet it does describe a contributory process that involved staff at multiple levels across regions to evaluate the AI system. Their study reports several areas where quantitative evaluation methods missed potentially dangerous failures that were only detected with observations from front-line staff (169).

As pressures grow to adopt AI systems throughout society, scientists will increasingly be asked to serve as arbiters of technology safety (135) across all stages in a system's life-cycle, from design and deployment to regulation and litigation (170, 171). Because AI systems are fundamentally sociotechnical in nature, their makers already face challenges of reliability, transparency, performance in context, and security at every level they operate in. In this article, we have argued that scientists can better address those sociotechnical challenges by collaborating with affected stakeholders in many parts of the research endeavor.

Public participation is not an automatic guarantor of better science. While AI has relied for decades on contributions of data from the public, many of those projects have been plagued with problems of bias that have resulted in serious flaws in the training and evaluation of AI systems (5, 128). Indeed, some of the most notorious AI failures have occurred when designers made their systems available on the open Internet for training (172). In this paper, we summarize concerns about research priorities, the problem of subjectivity, reductions in reliability, and the difficulties of cost, pace, and scale that effective participatory projects must account for.

Yet when undertaken with the level of care and expertise that is required for all reliable science, public involvement can improve the science of AI. Equipoise can bring together stakeholders in ways that improve the evidence through the scrutiny of disagreement. Participatory methods improve the precision and reliability of measurement by incorporating lived experience into the constructs that scientists seek to evaluate. Public involvement improves inference by strengthening the realism of assumptions in evaluation models. People with direct experience of AI systems can help explain how those systems behave in everyday circumstances that scientists may not be able to imagine. Finally, the public can serve as creative partners in the scientific and policy interpretation of evaluation findings. These evaluations may not resolve disputes entirely, but they can advance science and enable negotiations to be based on a more accurate, reliable understanding of AI systems in use.

Data, Materials, and Software Availability. There are no data underlying this work.

ACKNOWLEDGMENTS. We are grateful to all the community partners who we have learned from in the endeavor of participatory science. We also appreciate Tim Davies and Marianne Aubin Le Quéré for feedback. This project was supported by grants from the Templeton World Charity Foundation (https://doi.org/10.54224/33008), the John D. and Catherine T. MacArthur Foundation, the Heising-Simons Foundation, and the Siegel Family Endowment.

1. C. Ho, *Kaiser Nurses Protest Use of AI That They Say Could Put Patient Safety at Risk* (San Francisco Chronicle, 2024).
2. W. H. O. Guidance, *Ethics and Governance of Artificial Intelligence for Health* (World Health Organization, 2021).
3. L. Wright *et al*., "Null compliance: NYC local law 144 and the challenges of algorithm accountability" in *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (ACM, 2024), pp. 1701–1713.
4. B. Siwicki, *Kaiser Permanente's New Head of AI on 'Two Fundamental Shifts' the Technology will Enable* (Healthcare IT News, 2025).
5. A. Narayanan, S. Kapoor, A. I. Snake, *Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference* (Princeton University Press, 2024).
6. D. Cifci, G. P. Veldhuizen, S. Foersch, J. N. Kather, AI in computational pathology of Cancer: Improving diagnostic workflows and clinical outcomes? *Annu. Rev. Cancer Biol.* **7**, 57–71 (2023).
7. S. Rachuba, M. Reuter-Oppermann, C. Thielen, *Integrated Planning in Hospitals: A Review* (OR Spectroscopy, 2024).
8. V. Bellini *et al*., Artificial intelligence in operating room management. *J. Med. Syst.* **48**, 19 (2024).
9. D. P. Edelson *et al*., Early warning scores with and without artificial intelligence. *JAMA Netw. Open* **7**, e2438986 (2024).
10. S. A. Alowais *et al*., Revolutionizing healthcare: The role of artificial intelligence in clinical practice. *BMC Med. Educ.* **23**, 689 (2023).
11. W. Holmes, J. Persson, I. A. Chounta, B. Wasson, V. Dimitrova, *Artificial Intelligence and Education: A Critical View Through the Lens of Human Rights, Democracy and the Rule of Law* (Council of Europe, 2022).
12. S. Brayne, *Predict and Surveil: Data, Discretion, and the Future of Policing* (Oxford University Press, 2021).
13. M. T. Stevenson, *Assessing Risk Assessment in Action* (Minnesota Law Review, 2018).
14. I. Mathauer, M. Oranje, Machine learning in health financing: Benefits, risks and regulatory needs. *Bull. World Heal. Organ.* **102**, 216–224 (2024).
15. M. Hosseini, S. P. J. M. Horbach, Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Res. Integr. Peer Rev.* **8**, 4 (2023).
16. M. Kosmala, A. Wiggins, A. Swanson, B. Simmons, Assessing data quality in citizen science. *Front. Ecol. Environ.* **14**, 551–560 (2016).
17. E. Aceves-Bueno *et al*., The accuracy of citizen science data: A quantitative review. *Bull. Ecol. Soc. Am.* **98**, 278–290 (2017).
18. J. N. Matias, Humans and algorithms work together-so study them together. *Nature* **617**, 248–251 (2023).
19. A. Chouldechova *et al*., "AI red teaming through the lens of measurement theory" in *Neurips Safe Generative AI Workshop 2024* (Neural Information Processing Systems Foundation, 2024).
20. M. Young *et al*., *Participation Versus Scale: Tensions in the Practical Demands on Participatory AI* (First Monday, 2024).
21. M. S. Lam *et al*., End-user audits: A system empowering communities to lead large-scale investigations of harmful algorithmic behavior. *Proc. ACM Hum. Comput. Interact.* **6**, 1–34 (2022).
22. H. Shen, A. DeVos, M. Eslami, K. Holstein, Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proc. ACM Hum. Comput. Interact.* **5**, 1–29 (2021).
23. S. Lazar, A. Nelson, AI safety on whose terms? *Science* **381**, 138 (2023).
24. L. Weidinger *et al*., Sociotechnical safety evaluation of generative AI systems. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2310.11986 (Accessed 6 January 2025).
25. O. Elemento, C. Leslie, J. Lundin, G. Tourassi, Artificial intelligence in cancer research, diagnosis and therapy. *Nat. Rev. Cancer* **21**, 747–752 (2021).
26. P. A. Kulkarni, H. Singh, Artificial intelligence in clinical diagnosis: Opportunities, challenges, and hype. *JAMA* **330**, 317–318 (2023).
27. Y. Kumar, A. Koul, R. Singla, M. F. Ijaz, Artificial intelligence in disease diagnosis: A systematic literature review, synthesizing framework and future research agenda. *J. Ambient Intell. Humaniz. Comput.* **14**, 8459–8486 (2023).
28. E. Pierson, D. M. Cutler, J. Leskovec, S. Mullainathan, Z. Obermeyer, An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nat. Med.* **27**, 136–140 (2021).
29. G. C. Bowker, *Sorting Things Out: Classification and Its Consequences* (MIT press, 2000).
30. E. Neri, F. Coppola, V. Miele, C. Bibbolino, R. Grassi, Artificial intelligence: Who is responsible for the diagnosis? *La Radiol. Med.* **125**, 517–521 (2020).
31. A. Arora *et al*., The value of standards for health datasets in artificial intelligence-based applications. *Nat. Med.* **29**, 2929–2938 (2023).

32. S. Barocas, M. Hardt, A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities* (MIT press, 2023).
33. B. Laufer, S. Jain, A. F. Cooper, J. Kleinberg, H. Heidari, "Four years of FAccT: A reflexive, mixed-methods analysis of research contributions, shortcomings, and future prospects" in *2022 ACM Conference on Fairness, Accountability, and Transparency* (ACM, 2022), pp. 401–426.
34. L. Eckhouse, K. Lum, C. Conti-Cook, J. Ciccolini, Layers of bias: A unified approach for understanding problems with risk assessment. *Crim. Justice Behav.* **46**, 185–209 (2019).
35. Y. Yacoby, B. Green, C. .L. Griffin Jr., F. Doshi-Velez, "If it didn't happen, why would I change my decision?: How judges respond to counterfactual explanations for the public safety assessment" in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (AAAI Press, 2022), vol. 10, pp. 219–230.
36. D. Pruss, "Ghosting the machine: Judicial resistance to a recidivism risk assessment instrument" in *2023 ACM Conference on Fairness, Accountability, and Transparency* (ACM, 2023), pp. 312–323.
37. A. Fine, S. Marsh, Judicial leadership matters (yet again): The association between judge and public trust for artificial intelligence in courts. *Discov. Artif. Intell.* **4**, 44 (2024).
38. D. Vela *et al*., Temporal quality degradation in AI models. *Sci. Rep.* **12**, 11654 (2022).
39. M. Lipsky, *Street-Level Bureaucracy: Dilemmas of the Individual in Public Service* (Russell Sage Foundation, 2010).
40. J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, Human decisions and machine predictions. *Q. J. Econ.* **133**, 237–293 (2018).
41. H. F. Cheng *et al*., "How child welfare workers reduce racial disparities in algorithmic decisions" in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22* (Association for Computing Machinery, 2022).
42. B. Shemesh, E. Coughlan, T. Horton, *Tech to Save Time: How the NHS Can Realise the Benefits* (The Health Foundation, 2025).
43. B. C. Das, M. H. Amini, Y. Wu, Security and privacy challenges of large language models: A Survey. *ACM Comput. Surv.* **57**, 152:1–152:39 (2025).
44. T. H. McCoy, R. H. Perlis, Temporal trends and characteristics of reportable health data breaches, 2010–2017. *Jama* **320**, 1282–1284 (2018).
45. H. T. Neprash *et al*., Trends in Ransomware Attacks on US Hospitals, Clinics, and Other Health Care Delivery Organizations, 2016-2021. JAMA Health Forum 3, e224873 (2022).
46. J. Kenway, C. Francois, S. Costanza-Chock, I. D. Raji, J. Buolamwini, Bug bounties for algorithmic harms? (2022). https://www.ajl.org/bugs. Accessed 30 March 2025.
47. M. Feffer, A. Sinha, W. H. Deng, Z. C. Lipton, H. Heidari, "Red-Teaming for generative AI: Silver bullet or security theater?" in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (AAAI Press, 2024), vol. 7, pp. 421–437.
48. A. B. Gilmore *et al*., Defining and conceptualising the commercial determinants of health. *Lancet* **401**, 1194–1213 (2023).
49. R. N. Proctor, L. Schiebinger, *Agnotology: The Making and Unmaking of Ignorance* (Stanford University Press, 2008).
50. K. Popper, *The Open Society and Its Enemies* (Routledge, 1947).
51. Canadian Medical Association Journal, The "file drawer" phenomenon: Suppressing clinical evidence. *CMAJ Can. Med. Assoc. J.* **170**, 437 (2004).
52. L. Groves, J. Metcalf, A. Kennedy, B. Vecchione, A. Strait, "Auditing Work: Exploring the New York City algorithmic bias audit regime" in *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (ACM, 2024), pp. 1107–1120.
53. N. I. Silber, *Test and Protest: The Influence of Consumers Union* (Holmes & Meier, 1983).
54. T. Dietz, E. Ostrom, P. C. Stern, The struggle to govern the commons. *Science* **302**, 1907–1912 (2003).
55. J. L. Shirk *et al*., Public participation in scientific research: A framework for deliberate design. *Ecol. Soc.* **17**, e26269051 (2012).
56. D. Burns, S. M. Ospina, J. Howard, *The SAGE Handbook of Participatory Research and Inquiry* (SAGE Publications Ltd., 2021).
57. S. R. Arnstein, A ladder of citizen participation. *J. Am. Inst. Plann.* **35**, 216–224 (1969).
58. C. Adelman, Kurt Lewin and the origins of action research. *Educ. Action Res.* **1**, 7–24 (1993).
59. D. Schuler, A. Namioka, *Participatory Design: Principles and Practices* (CRC Press, 1993).
60. N. Wallerstein, B. Duran, J. G. Oetzel, M. Minkler, *Community-Based Participatory Research for Health: Advancing Social and Health Equity* (John Wiley & Sons, 2017).
61. A. Fung, Varieties of participation in complex governance. *Public Adm. Rev.* **66**, 66–75 (2006).
62. M. Reuchamps, J. Vrydagh, Y. Welp, Eds., *De Gruyter Handbook of Citizens' Assemblies* (De Gruyter, 2023).
63. E. G. Christopherson, D. A. Scheufele, B. Smith, *The Civic Science Imperative (SSIR)* (Stanford Social Innovation Review, 2018).
64. J. Garlick, P. Levine, Where civics meets science: Building science for the public good through Civic Science. *Oral Dis.* **23**, 692–696 (2017).
65. E. L. Boyer, The scholarship of engagement. *Bull. Am. Acad. Arts Sci.* **49**, 18–33 (1996).
66. M. Beaulieu, M. Breton, A. Brousselle, Conceptualizing 20 years of engaged scholarship: A scoping review. *PLoS One* **13**, e0193201 (2018).
67. P. Beresford, S. Croft, "User controlled research: Scoping review" (SSCR Scoping Review 5, London School for Social Care Research, 2012).
68. R. Putnam, Announcing a new name for this Association (2023). https://participatorysciences.org/2023/07/14/announcing-a-new-name-for-this-association/. Accessed 4 March 2025.
69. A. Halfaker, R. S. Geiger, ORES: Lowering barriers with participatory machine learning in Wikipedia. *Proc. ACM Hum. Comput. Interact.* **4**, 1–37 (2020).
70. J. Zong, J. N. Matias, Data refusal from below: A framework for understanding, evaluating, and envisioning refusal as design. *ACM J. Responsible Comput.* **1**, 1–23 (2024).
71. J. A. Goland, Algorithmic disgorgement: Destruction of artificial intelligence models as the FTC's newest enforcement tool for bad data. *Rich. JL Tech.* **29**, 1 (2022).
72. M. L. Gray, S. Suri, *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass* (Harper Business, 2019).
73. S. T. Roberts, *Behind the Screen* (Yale University Press, 2019).
74. M. Fuentes, B. M. Van Doren, D. Fink, D. Sheldon, BirdFlow: Learning seasonal bird movements from eBird data. *Methods Ecol. Evol.* **14**, 923–938 (2023).
75. C. J. Rosenblatt *et al*., Highly specialized recreationists contribute the most to the citizen science project eBird. *Ornithol. Appl.* **124**, duac008 (2022).
76. I. Shumailov *et al*., AI models collapse when trained on recursively generated data. *Nature* **631**, 755–759 (2024).
77. M. Ponti, A. Seredko, Human-machine-learning integration and task allocation in citizen science. *Humanit. Soc. Sci. Commun.* **9**, 1–15 (2022).
78. R. A. Hart, *Children's Participation: The Theory and Practice of Involving Young Citizens in Community Development and Environmental Care* (Routledge, 2013).
79. D. Kocman *et al*., Toolkit for conducting citizen science activities in environmental epidemiology. *Front. Environ. Sci.* **11**, 1177413 (2023).
80. V. Pandey *et al*., "Galileo: Citizen-led experimentation using a social computing system" in *Proceedings of the 2021 CHI conference on human factors in computing systems* (ACM, 2021), pp. 1–14.
81. R. Glennerster, K. Takavarasha, *Running Randomized Evaluations: A Practical Guide* (Princeton University Press, Princeton, NJ, 2014).
82. A. Birhane *et al*., "Power to the people? Opportunities and challenges for participatory AI" in *Equity and Access in Algorithms, Mechanisms, and Optimization* (ACM, 2022), pp. 1–8.
83. D. Zytko, P. J. Wisniewski, S. Guha, E. P. S. Baumer, M. K. Lee, "Participatory design of AI systems: Opportunities and challenges across diverse users, relationships, and application domains" in *CHI Conference on Human Factors in Computing Systems Extended Abstracts* (ACM, 2022), pp. 1–4.
84. A. Zhang *et al*., Deliberating with AI: Improving decision-making for the future through participatory AI design and stakeholder deliberation. *Proc. ACM Hum. Comput. Interact.* **7**, 1–32 (2023).
85. F. Delgado, S. Yang, M. Madaio, Q. Yang, "The participatory turn in AI design: Theoretical foundations and the current state of practice" in *Equity and Access in Algorithms, Mechanisms, and Optimization* (ACM, 2023), pp. 1–23.
86. H. Suresh *et al*., "Participation in the age of foundation models" in *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (ACM, 2024), pp. 1609–1621.
87. A. Duerinckx *et al*., Co-creating artificial intelligence: Designing and enhancing democratic AI solutions through citizen science. *Citiz. Sci. Theory Pract.* **9**, e732 (2024).
88. D. G. Robinson, *Voices in the Code: A Story About People, Their Values, and the Algorithm They Made* (Russell Sage Foundation, 2022).
89. S. Costanza-Chock, I. D. Raji, J. Buolamwini, "Who audits the auditors? Recommendations from a field scan of the algorithmic auditing ecosystem" in *2022 ACM Conference on Fairness, Accountability, and Transparency* (ACM, 2022), pp. 1571–1583.
90. W. H. Deng *et al*., "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* (ACM, 2023), pp. 1–18.
91. D. Reisman, J. Schultz, K. Crawford, M. Whittaker, Algorithmic impact assessments: A practical Framework for Public Agency (2018). https://ainowinstitute.org/publications/algorithmic-impact-assessments-report-2. Accessed 3 May 2025.
92. J. Metcalf, E. Moss, E. A. Watkins, R. Singh, M. C. Elish, "Algorithmic impact assessments and accountability: The co-construction of impacts" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (ACM, 2021), pp. 735–746.
93. L. Lin *et al*., Against the Achilles' heel: A survey on red teaming for generative models. *J. Artif. Intell. Res.* **82**, 687–775 (2025).
94. T. Gillespie, R. Shaw, M. L. Gray, J. Suh, AI red-teaming is a sociotechnical system. arXiv [Preprint] (2024). https://doi.org/10.48550/arXiv.2412.09751 (Accessed 3 May 2025).
95. J. N. Matias, M. Mou, "CivilServant: Community-Led experiments in platform governance" in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (ACM, 2018), pp. 1–13.
96. B. Zhang *et al*., "Public opinion toward artificial intelligence" in *The Oxford Handbook of AI Governance*, J. B. Bullock, Ed. (Oxford University Press, 2024).
97. O. F. Reid, A. Colom, R. Modhvadia, "What do the public think about AI?" (2023). https://www.adalovelaceinstitute.org/evidence-review/what-do-the-public-think-about-ai/. Accessed 21 February 2025.
98. K. Crockett *et al*., "Building trust - The people's panel for AI" in *2023 IEEE Conference on Artificial Intelligence (CAI)* (ACM, 2023), pp. 168–170.
99. S. Atwood, K. Bozentko, "U.S. public assembly on high risk Artificial Intelligence (AI)" (2023). https://cndp.us/ai. Accessed 3 May 2025.
100. T. Davies, "Involving the public in AI policymaking - Experiences from the People's Panel on AI" (2024). https://connectedbydata.org/projects/2023-peoples-panel-on-ai. Accessed 3 May 2025.
101. J. Stilgoe, AI has a democracy problem. Citizens' assemblies can help. *Science* **385**, adr6713 (2024).
102. S. Purpura, C. Cardie, J. Simons, "Active learning for E-rulemaking: public comment categorization" in *Proceedings of the 2008 International Conference on Digital Government Research, dg.o '08* (Digital Government Society of North America, 2008), pp. 234–243.
103. A. Press, How the U.S. labor movement is confronting AI. *New Labor Forum* **33**, 15–22 (2024).
104. L. C. Irani, M. S. Silberman, "Turkopticon: Interrupting worker invisibility in amazon mechanical turk" in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (ACM, 2013), pp. 611–620.

105. N. Jones *et al*., Lived experience, research leadership, and the transformation of mental health services: Building a researcher pipeline. *Psychiatry Serv.* **72**, 591–593 (2021).
106. C. Okoroji, T. Mackay, D. Robotham, D. Beckford, V. Pinfold, Epistemic injustice and mental health research: A pragmatic approach to working with lived experience expertise. *Front. Psychiatry* **14**, e1114725 (2023).
107. J. N. Matias, A. Hounsel, N. Feamster, Software-supported audits of decision-making systems: Testing Google and Facebook's political advertising policies. *Proc. ACM Hum. Comput. Interact.* **6**, 1–19 (2022).
108. E. Vázquez, M. Kim, M. E. Santaella, Lived experience experts: A name created by us for us. *Expert. Rev. Hematol.* **16**, 7–11 (2023).
109. M. Murphy, "Feminism, surveys, and toxic details" in *Sick Building Syndrome and the Problem of Uncertainty: Environmental Politics, Technoscience, and Women Workers* (Duke University Press, 2006).
110. P. B. English, M. J. Richardson, C. Garzón-Galvis, From crowdsourcing to extreme citizen science: Participatory research for environmental health. *Annu. Rev. Public Heal.* **39**, 335–350 (2018).
111. N. Weisshaupt, T. Lehtiniemi, J. Koistinen, Combining citizen science and weather radar data to study large-scale bird movements. *Ibis* **163**, 728–736 (2021).
112. P. Gill, C. Diot, L. Y. Ohlsen, M. Mathis, S. Soltesz, User initiated Internet data for the research community. ACM SIGCOMM CCR **52**, 34–37 (2022).
113. B. M. Viola, K. J. Sorrell, R. H. Clarke, S. P. Corney, P. M. Vaughan, Amateurs can be experts: A new perspective on collaborations with citizen scientists. *Biol. Conserv.* **274**, 109739 (2022).
114. A. Irwin, No PhDs needed: How citizen science is transforming research. *Nature* **562**, 480–482 (2018).
115. H. Riesch, C. Potter, Citizen science as seen by scientists: Methodological, epistemological and ethical dimensions. *Public Underst. Sci.* **23**, 107–120 (2014).
116. D. E. Stokes, *Pasteur's Quadrant: Basic Science and Technological Innovation* (Brookings Institution Press, 2011).
117. N. Laberge *et al*., Subfield prestige and gender inequality among U.S. computing faculty. *Commun. ACM* **65**, 46–55 (2022).
118. K. Bodker, F. Kensing, J. Simonsen, *Participatory IT Design: Designing for Business and Workplace Realities* (MIT press, 2009).
119. A. Gantman *et al*., A pragmatist philosophy of psychological science and its implications for replication. *Behav. Brain Sci.* **41**, e127 (2018).
120. K. Lewin, Psychology and the process of group living. *J. Soc. Psychol.* **17**, 113–131 (1943).
121. R. B. Cialdini, Full-cycle social psychology. *Appl. Soc. Psychol. Annu.* **1**, 21–47 (1980).
122. C. R. Mortensen, R. B. Cialdini, Full-cycle social psychology for theory and application: Full-cycle social psychology. *Soc. Pers. Psychol. Compass* **4**, 53–63 (2010).
123. S. Bødker, C. Dindler, O. S. Iversen, R. C. Smith, *Participatory Design* (Springer Nature, 2022).
124. S. Harding, After Mr. Nowhere: What kind of proper self for a scientist? *Fem. Philos. Q.* **1**, e2 (2015).
125. R. Proctor, *Value-Free Science?: Purity and Power in Modern Knowledge* (Harvard University Press, 1991).
126. L. Von Ahn, "Human computation" in *Proceedings of the 4th international conference on Knowledge capture* (ACM, 2007), pp. 5–6.
127. S. U. Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York University Press, 2020).
128. K. Yang, K. Qinami, L. Fei-Fei, J. Deng, O. Russakovsky, "Towards fairer datasets: filtering and balancing the distribution of the people subtree in the ImageNet hierarchy" in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20* (Association for Computing Machinery, 2020), pp. 547–558.
129. A. Birhane, V. U. Prabhu, "Large image datasets: A pyrrhic win for computer vision?" in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)* (IEEE, 2021), pp. 1536–1546.
130. S. Ghai, L. Fassi, F. Awadh, A. Orben, Lack of sample diversity in research on adolescent depression and social media use: A scoping review and meta-analysis. *Clin. Psychol. Sci.* **11**, 759–772 (2023).
131. R. Bonney *et al*., Citizen science: A developing tool for expanding science knowledge and scientific literacy. *BioScience* **59**, 977–984 (2009).
132. T. J. Bird *et al*., Statistical solutions for error and bias in global citizen science datasets. *Biol. Conserv.* **173**, 144–154 (2014).
133. O. J. Robinson, V. Ruiz-Gutierrez, D. Fink, Correcting for bias in distribution modelling for rare species using citizen science data. *Divers. Distrib.* **24**, 460–472 (2018).
134. A. I. López, M. I. F. Mójer, Rethinking civic science funding to better support community engagement. *J. Sci. Policy Gov.* **23**, jspg230204 (2024).
135. A. Orben, J. N. Matias, Fixing the science of digital technology harms. *Science* **388**, 152–155 (2025).
136. K. Popper, *The Logic of Scientific Discovery* (Routledge, 2005).
137. S. P. Hey, A. J. London, C. Weijer, A. Rid, F. Miller, Is the concept of clinical equipoise still relevant to research? *BMJ* **359**, j5787 (2017).
138. B. Freedman, Equipoise and the ethics of clinical research. *New Engl. J. Med.* **317**, 141–145 (1987).
139. M. J. Nadworny, *Scientific Management and the Unions, 1900–1932; a Historical Analysis* (Harvard University Press, 1955).
140. D. MacKay, The ethics of public policy RCTs: The principle of policy equipoise. *Bioethics* **32**, 59–67 (2018).
141. R. Kohavi, R. M. Henne, D. Sommerfield, "Practical guide to controlled experiments on the web: listen to your customers not to the hippo" in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM, 2007), pp. 959–967.
142. P. Amado *et al*., "Methodological report of the JEP-CEV-HRDAG Joint Project on data integration and statistical estimation" (2025). https://hrdag.org/CEV-JEP/20250306-methodological-report-EN.pdf. Accessed 3 May 2025.
143. D. Friedens-Warte, Guatemala – memory of silence: Report of the commission for historical clarification: Conclusions and recommendations. *Die Friedens-Warte* **74**, 511–547 (1999).
144. W. J. Clinton, Remarks by the president in roundtable discussions on peace efforts in Guatemala City (1999). https://www.presidency.ucsb.edu/node/229204. Accessed 14 August 2025.
145. J. E. Martinez, Analytic racecraft: Race-based averages create illusory group differences in perceptions of racism. *J. Exp. Psychol. Gen.* **153**, 3042–3061 (2024).
146. M. Cikara, J. E. Martinez, N. A. Lewis, Moving beyond social categories by incorporating context in social psychological theory. *Nat. Rev. Psychol.* **1**, 537–549 (2022).
147. I Institute, Beneath the surface (2021). https://invisible.institute/beneath-the-surface. Accessed 14 August 2025.
148. T. Reynolds-Tyler, T. Shah, "FAccT'23 keynote: The community built a model: Using participatory AI to analyze Chicago police data" (video recording, 2023). https://www.youtube.com/watch?v=kbPXUq-sVpQ. Accessed 14 August 2025.
149. T. U. Hauser, V. Skvortsova, M. D. Choudhury, N. Koutsouleris, The promise of a model-based psychiatry: Building computational models of mental ill health. *Lancet Digit. Heal.* **4**, e816–e828 (2022).
150. C. M. Gillan, R. B. Rutledge, Smartphones and the neuroscience of mental health. *Annu. Rev. Neurosci.* **44**, 129–151 (2021).
151. M. L. Gordon *et al*., "Jury learning: Integrating dissenting voices into machine learning models" in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (ACM, 2022), pp. 1–19.
152. C. Rudin *et al*., Interpretable machine learning: Fundamental principles and 10 grand challenges. *Stat. Surv.* **16**, 1–85 (2022).
153. W. J. Brady, K. McLoughlin, T. N. Doan, M. J. Crockett, How social learning amplifies moral outrage expression in online social networks. *Sci. Adv.* **7**, eabe5641 (2021).
154. R. Caplan, D. Boyd, Isomorphism through algorithms: Institutional dependencies in the case of Facebook. *Big Data Soc.* **5**, 2053951718757253 (2018).
155. M. Eslami *et al*., "First I "like" it, then I hide it: Folk theories of social feeds" in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (ACM, 2016), pp. 2371–2382.
156. J. N. Matias, Influencing recommendation algorithms to reduce the spread of unreliable news by encouraging humans to fact-check articles, in a field experiment. *Sci. Rep.* **13**, 11715 (2023).
157. D. Martin Jr., V. Prabhakaran, J. Kuhlberg, A. Smart, W. S. Isaac, Participatory problem formulation for fairer machine learning through community based system dynamics. arXiv [Preprint] (2020). https://doi.org/10.48550/arXiv.2005.07572 (Accessed 6 January 2025).
158. R. Vaithianathan, E. Putnam-Hornstein, N. Jiang, P, Nand, T. Maloney, *Developing Predictive Models to Support Child Maltreatment Hotline Screening Decisions: Allegheny County Methodology and Implementation* (Centre for Social Data Analytics, 2017).
159. S. Ho, G. Burke, *Here's How an AI Tool May Flag Parents with Disabilities* (APNews, 2023).
160. N. Kallus, A. Zhou, The fairness of risk scores beyond classification: Bipartite ranking and the xAUC metric. *Adv. Neural Inf. Process. Syst.* **32**, e05826 (2019).
161. M. Gerchick *et al*., "The devil is in the details: Interrogating values embedded in the Allegheny family screening tool" in *The Allegheny Family Screening Tool in 2023 ACM Conference on Fairness, Accountability, and Transparency* (ACM, 2023), pp. 1292–1310.
162. V. Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St. Martin's Press, 2018).
163. A. Deaton, Instruments, randomization, and learning about development. *J. Econ. Lit.* **48**, 424–455 (2010).
164. J. N. Matias, Going Dark, "Social factors in collective action against platform operators in the reddit blackout" in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (ACM, 2016), pp. 1138–1151.
165. H. H. Thorp, V. Vinson, Context matters in social media. *Science* **385**, 1393 (2024).
166. R. Kitchin, "Thinking critically about and researching algorithms" in *The Social Power of Algorithms*, D. Beer, Ed. (Routledge, 2019), pp. 14–29.
167. D. T. Campbell, Reforms as experiments. *Am. Psychologist* **24**, 409 (1969).
168. C. N. Association, California nurses association demand patient safeguards against artificial intelligence technology (2024). https://www.nationalnursesunited.org/press/cna-demand-patient-safeguards-against-artificial-intelligence. Accessed 6 January 2025.
169. C. H. Cain *et al*., Quality assurance during the rapid implementation of an AI-assisted clinical documentation support tool. *NEJM AI* **2**, AIcs2400977 (2025).
170. J. N. Matias, J. Penney, Science and causality in technology litigation. *J. Trust. Saf.* **2**, e233 (2025).
171. J. Metcalf, R. Singh, E. Moss, E. Tafesse, E. A. Watkins, "Taking algorithms to courts: A relational approach to algorithmic accountability" in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (ACM, 2023), pp. 1450–1462.
172. M. J. Wolf, K. Miller, F. S. Grodzinsky, Why we should have seen that coming: Comments on Microsoft's tay "experiment," and wider implications. *SIGCAS Comput. Soc.* **47**, 54–64 (2017).