JANUARY 2025

Assessing AI

Surveying the Spectrum of Approaches to Understanding and Auditing AI Systems

Miranda Bogen

With Contributions From
Chinmay Deshpande
Ruchika Joshi
Evani Radiya-Dixit
Amy Winecoff
Kevin Bankston

CENTER FOR DEMOCRACY & TECHNOLOGY

The **Center for Democracy & Technology (CDT)** is the leading nonpartisan, nonprofit organization fighting to advance civil rights and civil liberties in the digital age. We shape technology policy, governance, and design with a focus on equity and democratic values. Established in 1994, CDT has been a trusted advocate for digital rights since the earliest days of the internet. The organization is headquartered in Washington, D.C. and has a Europe Office in Brussels, Belgium.

CDT's AI Governance Lab develops and promotes adoption of robust, technically-informed solutions for the effective regulation and governance of AI systems. The Lab provides public interest expertise in rapidly developing policy and technical conversations, to advance the interests of individuals whose lives and rights are impacted by AI.

A report from

JANUARY 2025

Assessing AI

Surveying the Spectrum of Approaches to Understanding and Auditing AI Systems

Miranda Bogen

With contributions from Chinmay Deshpande, Ruchika Joshi, Evani Radiya-Dixit, Amy Winecoff, and Kevin Bankston.

Illustration and print layout by Timothy Hoagland.

References in this report include original links as well as links archived and shortened by the Perma.cc service. The Perma.cc links also contain information on the date of retrieval and archive.

CENTER FOR DEMOCRACY & TECHNOLOGY

ES Executive Summary

The importance of a strong ecosystem of AI risk management and accountability has only increased in recent years, yet critical concepts like *auditing*, *impact assessment*, *red-teaming*, *evaluation*, and *assurance* are often used interchangeably — and risk losing their meaning without a stronger understanding of the specific goals that drive the underlying accountability exercise. Articulating and mapping these goals against policy proposals and practitioner actions can be helpful in tuning accountability practices to best suit their desired aims.

Goals of AI assessment and evaluation generally fall under the following categories:

- **Inform**: practices that can *facilitate an understanding* of a system's characteristics and risks
- **Evaluate**: practices that involve *assessing the adequacy* of a system, safeguards or practices
- **Communicate**: practices that help *make systems and their impacts legible* to relevant stakeholders
- **Change**: practices that support *incentivizing changes in actor behavior*

Understanding the **scope of inquiry**, or the breadth or specificity of questions posed by an assessment or evaluation, can be particularly useful in determining whether that activity is likely to

Exploratory	Structured	Focused	Specific
***Broad exploration* of possible harms and impacts of a system, generally informed but unbounded by a set of known risks.** Broad and unbounded exploration of harms and impacts can lead to the discovery of otherwise unforeseen issues, elicit reflection on prioritization of and investment in known issues, and ensure consequential harms are not overlooked.	**Consideration of a *set of harms and impacts within a defined taxonomy*.** When faced with an expansive and uncertain landscape of harms, orienting efforts around predefined frameworks and taxonomies can support a coherent understanding of priorities, set baselines to ensure foundational issues are addressed, and appropriately reflect consensus expectations.	**Evaluation of a *specific harm or impact* or *assessment against a procedural requirement*.** As particular consequences of AI impacts become clear, robust investigation of these specific harms, grounded in mixed-method approaches, can enable their effective management — including facilitating understanding of when established approaches may be insufficient.	**Analysis of a specific harm or impact using a *defined benchmark, metric, or requirement*.** Assessing a system against defined standards can be highly compelling in leading to specific actions (e.g., motivating voluntary risk mitigations, triggering required remediation efforts, gating market access, or activating other accountability measures).

Table 1. A spectrum of approaches to scoping inquiry into an AI system, from broadest to narrowest.

Low independence	Medium independence	High independence
Direct and privileged access to an organization or the technical systems it builds can enable thoughtful and thorough self-assessment that helps businesses proactively map, measure, and manage risk, and generates a documentation trail that can be used for further scrutiny or oversight.	Verification of system characteristics or business practices by a credible actor who is reasonably disinterested in the results of their assessment can motivate organizations to ensure their systems and processes meet expectations, generate confidence that reasonable efforts have been made to do so, and can trigger and inform additional scrutiny.	Impartial efforts to probe and validate the claims of systems and organizations — without constraint on the scope of inquiry or characterization of their findings — is necessary to surface relevant risks and to generate the necessary external pressure to ensure they are sufficiently prioritized.

Table 2. Differing degrees of independence when assessing an AI system, from least to most independent.

surface the most relevant impacts and motivate the desired actions. Scope of inquiry exists on a spectrum, but for ease of comprehension the following breakdown can be a useful mental model to understand different approaches and their respective theories of change.

Meanwhile, recognizing the **degree of independence** of particular assessment or evaluation efforts — for instance, whether developer or deployer of the system in question has control over the systems that will be included in a given inquiry, what questions may be asked about them, and whether and to what extent findings are disclosed — is important to understanding the degree of assurance such an effort is likely to confer.

Assessment and evaluation efforts can shift up and down each of these two axes somewhat independently: a low-specificity effort can be conducted in a high-independence manner, while a highly specific inquiry may be at the lowest level of independence and still lead to useful and actionable insights. Ultimately though, the ability of different efforts in driving desired outcomes relates to where they sit on this matrix.

Recommendations

- **Evaluation and assessment efforts should be scoped to best support a defined set of goals.** Practitioners and policymakers should be particularly attentive to whether the independence and/or specificity of their intended assessment and evaluation activities are well-matched to the goals they have for those efforts. While specificity can be effective in motivating action and driving accountability, practitioners and policymakers should not pursue specificity at the expense of broader inquiry.
- **Stakeholders involved in evaluation and assessment efforts should be transparent and clear about their goals, methods, and resulting recommendations or actions.** Auditors and assessors should clearly disclose the methods they have employed, any assumptions that shaped their work, and what version of a system was scrutinized. Evaluators should define the range of acceptable results or threshold that would pose a concern prior to conducting the test, and findings from lower-independence and higher-independence efforts should flow between internal and external actors to create constructive feedback loops.

- **Accountability efforts should include as broad an array of participants and methods as feasible, with sufficient resources to ensure they are conducted robustly.** AI assessment and evaluation activities must include a pluralistic set of approaches that are not constrained to practitioners with technical expertise but rather encompass a sociotechnical lens, (i.e., considering how AI systems might interact in unexpected ways with one another, with people, with other social or technical processes, and within their particular context of deployment). Robust audits and assessments require sufficient funding, time, personnel, and infrastructure, with compensation structures that support meaningful participatory approaches and higher independence efforts.

Ultimately, no one set of accountability actors, single scope of assessment, or particular degree of auditor independence can accomplish all of the goals that stakeholders have for AI assessment and evaluation activities. Instead, a constellation of efforts — from research, to assurance, to harm mitigation, to enforcement — will be needed to effectively surface and motivate attention to consequential impacts and harms on people and society.

Contents

Contents

01 Introduction

As society grapples with the rise of increasingly embedded and complex automated systems, the importance of a strong risk management and accountability ecosystem has only increased. Despite widespread agreement on the importance of making progress toward this goal, discussions have splintered across interconnected but sometimes vague concepts like *auditing*, *impact assessment*, *red-teaming*, *evaluation*, and *assurance*.

At some times, these concepts are used interchangeably, while in others they are quite distinct. For example, audits, impact assessment, and red-teaming have all been described as tools to identify risks from unsound systems (Casper et al., 2024; Storchan et al., 2024), but they may involve slightly different (though sometimes overlapping) methods. "Auditing" and "assurance" typically imply some degree of independent analysis (Radiya-Dixit & Neff, 2023), but what some call evaluation can involve nearly identical analytical techniques (Jones et al., 2024; Raji et al., 2023).

Researchers have devoted considerable attention to different facets of this conversation, and policymakers have listened, investing significant effort to make sense of this landscape and proposing reasonably concrete recommendations to translate discussion to action. Reports from the US National Telecommunication and Information Administration (NTIA) and the UK Department for Science, Innovation & Technology (DSIT) on AI accountability and AI assurance, respectively, have been valuable guideposts in an otherwise blurry landscape (DSIT, 2024; NTIA, 2024). Nevertheless, ambiguous terminology continues to fuel misaligned expectations among stakeholders and threatens to impair the implementation of these recommendations.

[A]mbiguous terminology continues to fuel misaligned expectations among stakeholders... This report aims to cut through this rhetorical fog and help stakeholders stay focused on our shared North Star: a well-developed and mature ecosystem [for AI assessment].

This report aims to cut through this rhetorical fog and help stakeholders stay focused on our shared North Star: a well-developed and mature ecosystem that properly incentivizes the identification and mitigation of AI's risks in order to judiciously harness its benefits. Many have tried to define what each approach *involves*, but these discussions can divert attention away from the important foundation: what these efforts are intended to *achieve*. This scaffolding is vital; once goals are clearly articulated, stakeholders can more effectively consider whether a given approach to evaluating risks is suitable for the intended purpose, and offer concrete recommendations for refining or enhancing different methods to meet those objectives.

The report proceeds as follows. First, we examine the array of desired outcomes that commonly motivate assessment and evaluation proposals and activities, synthesizing goals related to **informing** (facilitating an understanding of the characteristics of and risks posed by an AI system); **evaluating** (assessing the adequacy of a system, safeguards or practices); **communicating** (helping to make systems and their impacts legible to relevant stakeholders), and **changing** (incentivizing changes in actor behavior). Next, we discuss a number of illustrative statutes and policy proposals to demonstrate how they map against some of those desired outcomes. We then identify two primary dimensions

along which accountability activities appear to fall — **scope of inquiry** and **degree of independence** — to illuminate how different approaches relate to each other. Finally, we highlight conditions necessary for success within these dimensions as well as opportunities to adopt more holistic and inclusive methods across the ecosystem of approaches to inform effective implementation.

02 Goals of AI Assessment and Accountability Practices

Clearly articulating the intended goals of assessment and accountability practices is crucial to ensuring these efforts support their intended outcomes. Recognizing activities that sound similar but use different methods and yield divergent conclusions can help stakeholders understand why certain approaches may not be meeting particular goals. On the other hand, recognizing when different types of assessments with different terminologies nevertheless share similar objectives can make it easier for stakeholders to pinpoint and support the concrete practices most likely to achieve their desired outcomes.

Consider, for example, the concept of red teaming. Red teaming, which refers to adversarial testing of AI systems to elicit problematic outputs or vulnerabilities, is often invoked as a useful practice. However, red teaming exercises can have very different aims: in some cases, red teaming is intended to identify unforeseen harms, while in others it is presumed to be a means of assessing the adequacy of particular model safeguards. If the goals of a given red teaming exercise are not well-defined, organizations may fail to design these exercises in ways that are methodologically appropriate to achieving the intended aims. For instance, red teaming for harm discovery requires a broad and inclusive set of testers who are given free reign to interact with a system in realistic

scenarios, while red teaming to test model safeguards may benefit most from testers with specialized knowledge about the behavior they are aiming to elicit from a system, and perhaps even specific definitions and testing protocols to identify the prevalence or severity of a given harm or vulnerability.

Meanwhile, calls for the related activity of auditing might be motivated by the desire of some stakeholders for system developers to voluntarily demonstrate conformity with a specific and defined set of requirements (for instance, verifying that an enterprise has implemented a risk management system for high-impact AI systems as required by the EU AI Act), and of others for laying the groundwork for procedural challenges or regulatory enforcement (such as requiring AI systems in certain domains to proactively measure disparities in a system's outcomes that could later be challenged under civil rights laws). The former may require a voluntary set of guidelines and reasonably independent actors to evaluate technical systems or business practices against these standards, while the latter may need to be structured in a way that establishes a specific fact pattern or demonstrates that a system's outputs have exceeded a predefined threshold of concern. Each of these goals requires a different toolkit, and practitioners aiming to accomplish either will face different challenges along the way.

Table 3 synthesizes the goals that researchers, advocates, policymakers, and practitioners have explicitly and implicitly suggested can be supported by assessment and accountability practices. While there is some overlap between high level objectives, it can be helpful to group these goals can into four general areas:

- **Inform**: Practices that can facilitate an understanding of a system's characteristics and risks
- **Evaluate**: Practices that involve assessing the adequacy of a system, safeguards or practices
- **Communicate**: Practices that help make systems and their impacts legible to relevant stakeholders
- **Change**: Practices that support incentivizing changes in actor behavior

While many of these goals are shared by diverse stakeholders, some may be higher priorities for certain actors than others. For instance, commercial AI developers likely see many of these goals as supporting voluntary management of business risks. Public interest advocates, on the other hand, envision audits, impact assessments, and related activities as key tools for understanding the broader sociotechnical impacts of AI systems and for supporting robust enforcement by regulators to prevent or remediate AI harms. These goals span actors and phases across the AI development lifecycle and roughly map to existing frameworks like NIST's AI Risk Management Framework ("Map," "Measure," "Manage," and "Govern" – *see Table 3*). However, they are distinct enough that we offer this additional structure to help articulate desired goals with more specificity.

Nearly all of the goals appear to implicitly presume that the envisioned activities will somehow motivate institutions to mitigate the detected risks, but assumptions about *how* such mitigation and risk management will be incentivized differ dramatically. A clear understanding of these distinct theories of change will, ideally, help inform and support policy interventions that stakeholders are most confident will bring about the desired outcomes.

What goals have policymakers focused on?

Current regulations and policy proposals differ widely in their goals and methods. In this section, we assess a sample of proposed and enacted statutes to highlight some key recurring themes.

The proposed Validation and Evaluation for Trustworthy (VET) Artificial Intelligence Act, for instance, seeks to create the conditions where companies voluntarily undergo both internal and external assurance of artificial intelligence systems, with the goal of verifying claims regarding the functionality and testing of the AI system (VET AI Act, 2024). The proposal presupposes that such voluntary efforts will help ensure AI systems are fit for their intended purpose; anticipate errors or inconsistencies in testing, risk management, or internal governance; and identify vulnerabilities or negative societal impacts of the AI system.

High-level goal	Sub-goal	Mechanisms for assessment and accountability that support this goal
Inform	Identifying and understanding key characteristics or risks of system [MAP]	▪ *Map relevant characteristics* of a system (or organizational practices) to identify potential gaps and issues ▪ *Uncover what risks arise* from a system's general operation, unsound systems or practices, or system misuse (Casper et al., 2024) ▪ *Determine target characteristics or risks for further study, scrutiny or monitoring* (Casper et al., 2024) ▪ *Establish baselines* to guide iterative improvement (Weidinger, Barnhart, et al., 2024) ▪ *Identify directional gaps* in existing system safeguards or evaluation tools (Storchan et al., 2024) ▪ Inform development of *reliable and valid tests* to measure system behavior and impacts
	Assessing the magnitude or prevalence of these characteristics or risks to inform prioritization of development/remediation efforts [MEASURE]	▪ *Benchmark* model performance and risk against baselines and peers ▪ *Understand likelihood and magnitude of harms* in order to prioritize allocation of resources to further research harms and impacts and to develop effective mitigation methods ▪ *Understand relative effort and effectiveness* of different interventions and mitigations
Evaluate	Informing, motivating, or triggering changes in system design or intervention/mitigation [MANAGE]	▪ *Demonstrate existence or extent of issue* to motivate attention to and investment in remediation (Casper et al., 2024) ▪ *Inform recommendations for specific changes* to system design, implementation, or mitigations (UK Information Commissioner's Office, 2022) ▪ *Trigger policies* that require reduction of risk to a reasonable degree prior to further development or deployment, including implementation of auxiliary safeguards to address residual risk that primary mitigations cannot eliminate
	Investigating adequacy of safeguards or interventions [MEASURE]	▪ *Assess systems against predefined thresholds* in order to make decisions about mitigations or deployment (Weidinger, Barnhart, et al., 2024) ▪ *Evaluate sufficiency* of organizational processes or technical interventions in facilitating risk management efforts (Casper et al., 2024) ▪ *Ensure validity* of tests and benchmarks (Storchan et al., 2024) ▪ *Monitor effects of changes* and interventions to system characteristics, outcomes, and impacts over time
	Determining conformity against defined requirement(s) [GOVERN]	▪ Determine whether a system was developed *in line with legal requirements* ▪ Assess whether a system has *exceeded a defined threshold* (e.g., precision, outcome disparity, etc) ▪ Evaluate an organization's *practices against its claims* ▪ *Verify conformity* with procedural requirements (e.g., process-based standards)

Table 3. Goals of AI assessment and accountability practices.

High-level goal	Sub-goal	Mechanisms for assessment and accountability that support this goal
Communicate	Advancing broader awareness of relevant system details and impacts [GOVERN]	▪ Make results of assessments transparent and publicly accessible to *enhance legibility* of complex systems and their impacts (Groves, 2024) ▪ Share findings with regulators and researchers to *support further research on and broader understanding of* systems and their impacts (Casper et al., 2024) ▪ Use findings to help *facilitate deliberation and generate consensus* around how impacts should be defined and prioritized, and the appropriate methods to detect and remediate them (Moss et al., 2021)
	Demonstrating conformity with expectations [GOVERN]	▪ *Establish credibility* by documenting and disclosing adoption of expected practices ▪ *Demonstrate soundness of practices* by disclosing favorable results and/or effectiveness of remediation (Casper et al., 2024)
Change	Motivating conduct that meets procedural expectations and results in outcomes that fall within acceptable bounds [GOVERN]	▪ Enable users and customers to *make informed decisions about system adoption and use* based on assessment results ▪ *Gate market access based on verified conformity* with specific metrics or tests (Groves, 2024) ▪ *Create conditions for public scrutiny* or *impose monetary penalties* for subpar practices (Casper et al., 2024) ▪ *Require withdrawal or decommissioning* of systems that exceed defined thresholds ▪ Lay foundation to *justify legal remedy* for people who experience harmful impacts (Groves, 2024)
	Making the case that identified harms must be managed [GOVERN]	▪ *Stimulate public pressure* toward actors to address harms that have not been sufficiently attended to ▪ *Generate policy pressure* to stimulate additional investment in harm mitigation by highlighting gaps in existing incentive structures and governance tools ▪ *Motivate and/or attempt to justify imposition of penalties* for harms not previously enumerated
	Enabling challenge or enforcement action [GOVERN]	▪ *Generate evidence to establish standing* for legal or regulatory challenge ▪ *Provide evidence to support factual arguments* for legal or regulatory challenge ▪ *Justify imposition of penalties* by demonstrating relationship between system/actor and undesirable impact (Groves, 2024)

Table 3. (continued) Goals of AI assessment and accountability practices.

The EU AI Act takes a similar but more assertive approach, requiring that providers of high-risk AI systems demonstrate sound practices and undergo "conformity assessments" to determine whether they have implemented procedural requirements like effective systems for risk management, data governance, and transparency and human oversight (Artificial Intelligence Act, 2024). High-risk AI systems are also expected to be reviewed for appropriate levels of accuracy, robustness and cybersecurity throughout their lifecycle, in line with the goal of assessing the adequacy of safeguards (albeit via process-level checks).

[R]ecognizing when different types of assessments with different terminologies nevertheless share similar objectives can make it easier for stakeholders to pinpoint and support the concrete practices most likely to achieve their desired outcomes.

California's controversial bill SB 1047, or the Safe and Secure Innovation for Frontier Artificial Intelligence Models Act, would have directed developers of AI models created using more than a certain amount of computing power or training cost to assess those models for "critical harms" — defined as impacts leading to mass casualties, hundreds of millions of dollars of damages; reckless or negligent death, injury, or harm to property; or similarly harmful impacts — and to define procedures for mitigating those risks (Safe and Secure Innovation for Frontier Artificial Intelligence Models Act, 2024). The results of these tests would trigger the implementation of particular safeguards, up to and including model shutdown. The law would have required third-party auditors to assess developers' compliance with the statute, including the robustness of internal controls, and to identify areas for improvement. Audit reports were to be provided to the Attorney General on request, and the law would have empowered the AG to bring civil actions for violations of the statute. These components of the statute suggest motivations that included reducing risk, evaluating the sufficiency of organizational processes, and generating evidence to support regulatory challenges.

The EU Digital Services Act (DSA) likewise mandates annual independent audits of Very Large Online Platforms and Search Engines to drive changes in corporate conduct (Digital Services Act, 2022). Prior to these audits, audited parties are expected to provide descriptions of internal controls, historical data and benchmark metrics to measure performance, a preliminary risk

analysis, and unrestricted access to all data necessary for the audit. Auditors must then review and report on any inherent risks (risks of non-compliance inherent to the nature and application of the audited service), control risks (risks due to misstatements that the provider's internal controls failed to prevent or detect) and detection risks (risks due to the possibility of misstatements that remain undetected by the auditor). Based on the results, auditors must conclude the audit with one of three possible outcomes: "positive," "positive with comments," or "negative." Any report that is not entirely positive must include suggestions for improvement, along with a specified timeframe for implementation. These audits seemed to be envisioned to motivate companies to uncover risks and identify gaps in existing safeguards and tools, but in the absence of concrete expectations for how platforms should address specific risks, it will likely remain difficult to motivate appropriate interventions.

Colorado's Consumer Protections for Artificial Intelligence Act (SB 24-205) envisions impact assessments as a means to navigate the challenge of algorithmic discrimination, requiring deployers to describe intended use cases, potential risks, and bias measurement and mitigation strategies — and to convey that information to deployers (Consumer Protections for Artificial Intelligence Act, 2024). Based on a rebuttable presumption of "reasonable care," the bill appears to suggest that impact assessments, if conducted appropriately, can be effective in tackling algorithmic discrimination by informing and incentivizing changes in developer and deployer behavior.

Providing more guidance on how to calibrate assessments against defined benchmarks may be helpful in informing risk management activities and motivating sufficient investment in remediating identified issues. For example, the Algorithmic Accountability Act would have required entities using automated decision systems (ADS) or augmented critical decision processes (ACDP) to conduct impact assessments (Algorithmic Accountability Act, 2023). These assessments would have included requirements to evaluate the proposed system against previous decision-making processes, to document the purpose of the new system as well as any harms and benefits (informed by consultation with impacted communities), and to conduct ongoing testing of system performance across

relevant subpopulations. The law would have required covered entities to attempt to mitigate "likely material negative impact," indicating goals of compelling behavior change as well as creating the conditions for regulators to enforce against seemingly subpar risk management practices.

By more clearly mapping the goals that enacted and proposed policies seem to envision, stakeholders may be better positioned to identify and advocate for the necessary conditions for the success of these efforts, such as defining relevant methods and benchmarks and ensuring policymakers understand what their proposals are more and less likely to accomplish.

03 Dimensions of AI Assessment and Accountability

To understand whether and how different assessment activities are likely to support particular goals outlined in Table 3, we find it useful to situate such efforts in relation to one another along two primary dimensions: the **scope of inquiry**, or the specificity of the question being asked or hypotheses being tested, and the **degree of independence** of the examiner, or how much control the entity being tested has over the nature of the exercise and the framing of its outcomes. These dimensions are often conflated, but each axis can have a significant and distinct impact on the legibility of assessment outcomes to different stakeholders, the efficacy of findings in triggering different goals, and the level of confidence a process is likely to produce.

While each axis exists on a continuous spectrum, we divide them into several rough categories for ease of understanding. We start with an exploration of how the scope of inquiry reflects both different degrees of understanding of a given AI system as well as different theories of change. We then overlay the dimension of independence, which has meaningful implications both for the level of scrutiny allowed in an assessment and for the confidence its results inspire.

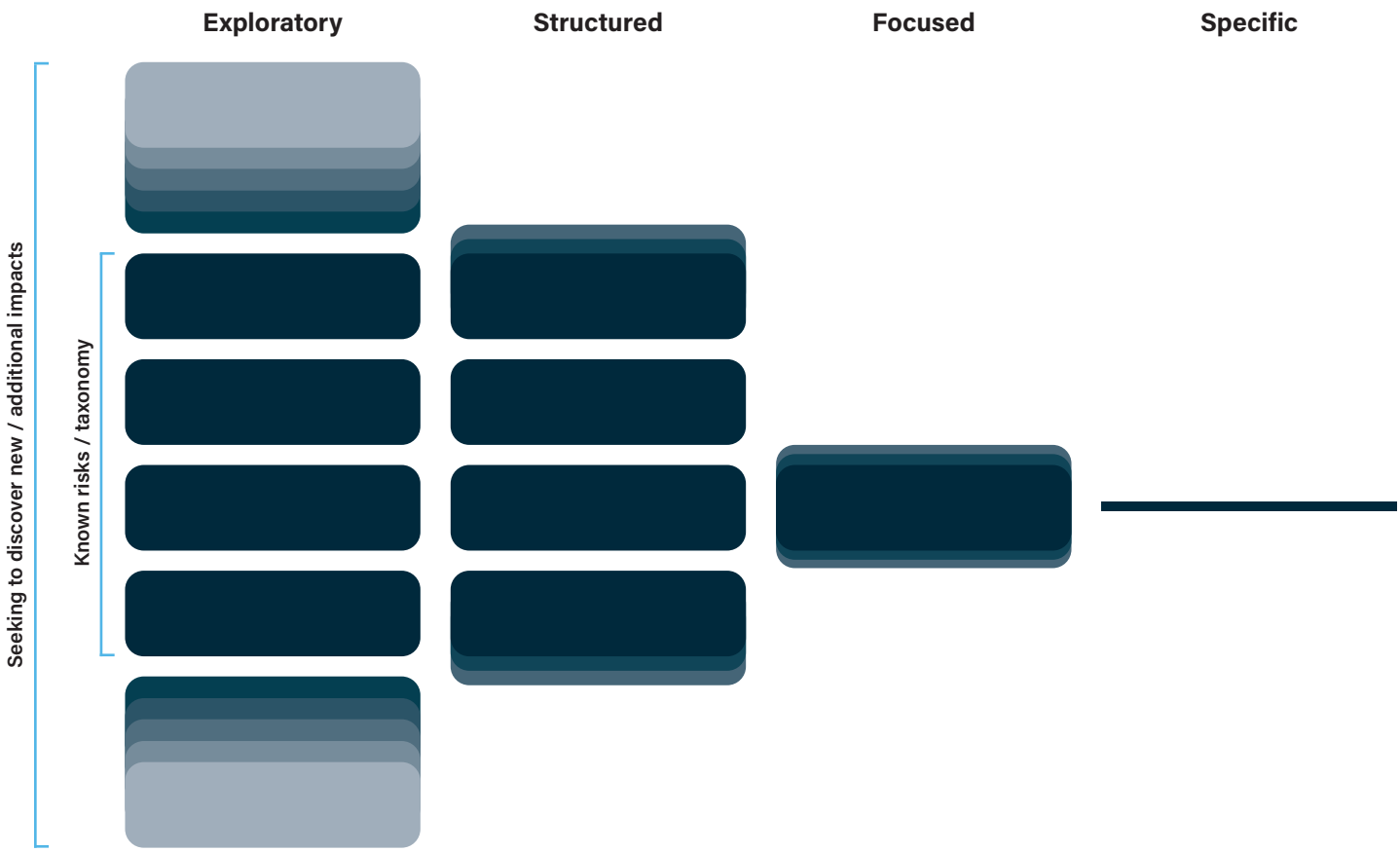

Figure 1. The spectrum of scope in AI assessments, from broadest to narrowest.

Scope of Inquiry

Researchers have noted the need to balance the benefits of precise objectives for audit- and assessment-related exercises with the importance of understanding broader contexts and implications that narrow investigations are prone to miss (Birhane et al., 2024; Burt & Leong, 2024). Some of the goals identified above will be served best by substantially broader exploration of harms and impacts, while others demand specificity enabled by precisely-defined hypotheses and methodologies.

The breadth of inquiry into AI systems can be represented as a spectrum, from entirely undefined and exploratory exercises to measurements of a specific metric. For convenience, we segment this spectrum into four broad categories: **exploratory**, **structured**, **focused**, and **specific**, starting with the widest scope and proceeding to the narrowest to reflect the logical process that practitioners and stakeholders may go through when envisioning and developing a new AI system. (As certain AI systems and

	Exploratory	Structured	Focused	Specific
Description	***Broad exploration* of possible harms and impacts of a system, generally informed but unbounded by a set of known risks.**	**Consideration of a *set of harms and impacts within a defined taxonomy*.**	**Evaluation of a *specific harm or impact* or *assessment against a procedural requirement*.**	**Analysis of a specific harm or impact using a *defined benchmark, metric, or requirement*.**
Theory of change	Broad and unbounded exploration of harms and impacts can lead to the discovery of otherwise unforeseen issues, elicit reflection on prioritization of and investment in known issues, and ensure consequential harms are not overlooked.	When faced with an expansive and uncertain landscape of harms, orienting efforts around predefined frameworks and taxonomies can support a coherent understanding of priorities, set baselines to ensure foundational issues are addressed, and appropriately reflect consensus expectations.	As particular consequences of AI impacts become clear, robust investigation of these specific harms, grounded in mixed-method approaches, can enable their effective management — including facilitating understanding of when established approaches may be insufficient.	Assessing a system against defined standards can be highly compelling in leading to specific actions (e.g., motivating voluntary risk mitigations, triggering required remediation efforts, gating market access, or activating other accountability measures).
Examples	▪ *Open-ended red-teaming to identify unforeseen harms* ▪ *Interviews with potentially impacted communities* ▪ *Broad evaluation of organizational practices*	▪ *Human rights impact assessments* ▪ *Assessing rights and safety impacts of a system* ▪ *Evaluation against harm taxonomies like those found in the NIST Generative AI Profile*	▪ *Bias audits* ▪ *Exploration of a system's impact on elections* ▪ *Investigating a system's contribution to non-consensual intimate imager (NCII)* ▪ *EU AI Act conformity assessment*	▪ *Measuring a system's adverse impact ratio* ▪ *Conducting a specific technical evaluation of a model's toxicity* ▪ *Verifying a company's assertion about data practices*

Table 4. A spectrum of approaches to scoping inquiry into an AI system, from broadest to narrowest.

assessments become more familiar and their risks more readily apparent, assessment and evaluation efforts may adopt narrower scopes of inquiry — but practitioners should be cautious about jumping prematurely to specific metrics or definitions of impact that can obscure important concerns.)

Activities conducted within each layer of specificity tend to be motivated by different theories of change, which are often unstated. Lack of clarity about these theories of change makes it harder to have constructive discussions about their utility.

In the following sections, we unpack when each type of assessment may be most useful, highlight key risks and limitations, and importantly, recommend approaches to make each type of effort more likely to achieve its goals.

Exploratory

Broad exploration of possible harms and impacts of a system, generally informed but unbounded by a set of known risks.

Theory of change: broad and unbounded exploration of harms and impacts will lead to the discovery of otherwise unforeseen issues, elicit reflection on prioritization of and investment in known issues, and help ensure consequential harms are not overlooked.

Experts have highlighted the importance of assessing AI systems' impacts broadly, holistically, and in an open-ended manner so that the widest range of potential harms can be identified and mitigated (Metcalf et al., 2021; Raji et al., 2023). Such inquiries have been described as tools for developers to reflect on the systems they are building (Raji et al., 2023), to discover otherwise unforeseen harms and unknown unknowns (Selbst, 2021), to advance more complete understanding of human and societal impacts (Anthropic, 2024a), and to inform prioritization of risks (Groves, 2022). This sort of **exploratory** inquiry has been cited as being particularly important for increasingly general-purpose technologies whose downstream use and behavior remain deeply uncertain (Anthropic, 2024a; Burt & Leong, 2024).

While exploration of harms within this paradigm tends to be grounded in a set of familiar issues, a key feature of efforts in this category is that exploration is not constrained to existing methods or taxonomies, allowing for richer understanding of the relevant impacts of AI systems (Metcalf et al., 2021). Proponents of a more exploratory approach argue that overly specifying analysis of a system risks reinforcing a narrow understanding of what risks are most pertinent or urgent (Groves, 2022), generating coarse and unreliable conclusions (Selbst, 2021), and overlooking a system's societal or longer-term impacts (Solaiman et al., 2024).

Exploratory assessment in practice

Exploratory inquiry has been used by developers of AI systems and external actors to proactively spot failure modes and to challenge overly narrow assertions of impact (Raji et al., 2020). For instance, NIST's ARIA (Assessing Risks and Impacts of AI) project aims to consider both pre-specified and unspecified risks and impacts of large language models to help build tools, measurement methods and metrics to support risk assessment (NIST, n.d.). Red-teamers for OpenAI's GPT-4o were instructed to "carry out exploratory capability discovery" and "assess novel potential risks posed by the model" in addition to stress testing interventions intended to address previously identified risks (OpenAI, 2024). A collaborative effort between developers of an AI-powered healthcare system and social science researchers involved an open-ended study of the human and technical implications of integrating that system into a hospital setting, illuminating the importance of nursing staff in interpreting and communicating the recommendations of the system (Sendak et al., 2020). And other stakeholders have used broadly-scoped methods like grassroots activism and humanitarian documentation efforts to surface relevant harms of AI-powered systems and motivate intervention (Birhane et al., 2024).

Opportunities and limitations

Exploratory approaches may be most useful in two circumstances: cases of novel or general purpose technologies in which an understanding of potential harms remains underdeveloped, and cases where methods to define targets of assessment and accountability have prematurely crystallized in a way that neglects important impacts and disempowers affected communities (Anthropic, 2024a; Metcalf et al., 2021). In both cases, surfacing a variety of impacts using a pluralistic array of methods can

motivate attention to remediating those issues. Understanding the fullest landscape of impacts is necessary to ensure appropriate prioritization among those impacts and to foster broader awareness of potential (and actual) harms (Vecchione et al., 2021). And inquiries unconstrained by predefined taxonomies or methods can surface examples of harmful impacts that might otherwise be missed.

Example goals that may benefit from this approach:

Inform

- Map relevant characteristics of a system or organization to identify potential gaps and issues
- Uncover what risks arise from a system's general operation, unsound systems or practices, or system misuse
- Determine targets for further study, scrutiny or monitoring

Communicate

- Use findings to help facilitate deliberation and generate consensus around how impacts should be defined and prioritized, and the appropriate methods to detect and remediate them
- Stimulate public pressure toward actors to address harms that have not been sufficiently attended to
- Generate policy pressure to stimulate additional harm mitigation by highlighting gaps in existing incentive structures and governance tools
- Generate evidence to establish standing for legal or regulatory challenge

At the same time, exploratory approaches tend to be challenged by claims that such efforts lack rigor — for example, not employing sound statistical sampling or experimental controls (Vecchione et al., 2021). Without standard methods or defined approaches, it can be harder to assess the adequacy of exploratory efforts or defend the conclusiveness of their results (Groves, 2022). If evaluators lack sufficient expertise, time, resources, methodological tools, or access to lived experience to understand a system's potential impacts, they may overlook the very issues such an exercise is intended to discover (Galindo et al., 2024; Groves, 2022; Weidinger, Mellor, et al., 2024). The method's inherent flexibility means that institutions conducting exploratory assessments have tremendous latitude to define and navigate the landscape of risks as they see fit, avoiding or downplaying topics that may implicate their products or services most directly. To be done well, open-ended assessments can be resource-intensive, which in some cases may divert attention or resources from investigating or mitigating already identified risks. If these dynamics are not recognized or addressed, the resulting incomplete assessment exercises may be used to justify proceeding with development or deployment, or leveraged to unduly bolster an organization's credibility (Groves, 2022). Finally, exploratory inquiries present natural opportunities to engage with external stakeholders and impacted communities, but insufficient capacity, investment, or willingness to incorporate findings can lead to tokenism and participation washing (Groves, 2022; Moss et al., 2021).

Considerations that can advance a more holistic approach

Exploratory inquiries into the impacts of AI systems can be conducted by internal or external actors. While internal actors like developers or first-party auditors may have privileged access to a system and its surrounding context that can help them envision harms prior to launch or that may be challenging to observe using adversarial methods, the most robust exploratory efforts will consider a broader definition of impacts, informed by a wide array of stakeholders. Therefore, based on recommendations drawn from (Groves, 2022; Moss et al., 2021; Radiya-Dixit, 2025; Storchan et al., 2024; Vecchione et al., 2021; Weidinger et al., 2023; Weidinger, Mellor, et al., 2024), assessors should:

- Prioritize building internal teams that are comprised of staff with domain expertise — and where possible, diverse lived experience — in order to conduct robust preliminary exploration of impacts.
- Expansively consider impacts of systems at the technical, human, and societal level, taking into account immediate impacts, harms that may be cumulative over time, and systemic impacts from complex system interactions or second-order effects.
- Establish a variety of mechanisms to solicit feedback and input (e.g., written comments, participatory workshops, user feedback, crowdsourcing, bug bounties, public red-teaming) and employ ethnographic methods to generate deep, qualitative insights about a system's potential impacts.
- Allocate sufficient time and resources to conduct stakeholder consultations, including recruiting a demographically diverse range of participants and devoting effort to explaining systems and their anticipated impacts in understandable language. Where directly affected individuals may face challenges participating, include organizations who advocate on behalf of those communities.
- Build sustained opportunities for engagement, such as multiple touchpoints for feedback, standing groups for regular consultation, or organic channels for stakeholders to provide input and feedback as issues emerge.
- Reduce barriers to participatory activities, such as providing transportation, interpretation, translation, compensation, and childcare.
- Pursue commitments to incorporate feedback from public consultations into the development or updating of the system(s) in question, even if doing so may appear to diverge from the developer or deployer's immediate interests. Ideally, organizations would build opportunities for co-design such that stakeholders are empowered to directly impact decisions about the system in question. At minimum, assessors can follow up with consulted communities to share what changes have been made in response to their input.
- Recognize and compensate those consulted in stakeholder engagements to avoid extractive or exploitative dynamics.

Consultation and compensation should be structured to avoid actual or perceived conflicts of interest (e.g., consider avoiding non-disclosure agreements where they are not necessary or overly filtering participant feedback or findings through corporate channels).

Structured

Consideration of a set of harms and impacts within a defined taxonomy.

Theory of change: When faced with an expansive and uncertain landscape of harms, orienting efforts around predefined frameworks and taxonomies supports a coherent understanding of priorities, sets baselines to ensure foundational issues are addressed, and appropriately reflects consensus expectations.

As any researcher tasked with exploring an undefined problem space knows, aimless inquiry into a new domain – although potentially helpful in uncovering previously unconsidered risks – is less likely to yield meaningful and actionable insights than efforts that are informed by and build on existing observations and expertise. For AI developers and accountability actors tasked with identifying and remediating harms, working from existing taxonomies can help teams more quickly understand relevant societal contexts, avoid duplicative analysis, facilitate structured decision making about prioritization, and organize what may otherwise seem like a daunting undertaking. **Structured** analysis — or the consideration of a set of harms and impacts within a defined taxonomy — can be based on one or more internal or external taxonomies, a set of laws or regulations, a prioritized list of issues surfaced by a more exploratory analysis, or an organization's articulated policies or practices (Galdon Clavell, 2024; Raji et al., 2020; Storchan et al., 2024).

Structured assessment in practice

Fundamental rights impact assessments (FRIAs) are required by the EU AI Act for high risk AI systems. These assessments are used to compare harms likely posed by a system to rights enumerated in the European Union Charter on Fundamental Rights (Waem et al., 2024). Human rights impact assessments (HRIAs) have similarly been cited as useful in systematically considering systems' or business practices' impact on categories of rights articulated in the Universal Declaration of Human Rights. As part of implementation of President Biden's Executive Order on AI, federal agencies were tasked with identifying whether AI systems fell into predefined "rights-impacting" or "safety-impacting" domains (S. D. Young, 2024) . And a prominent public red-teaming exercise at the DEFCON conference structured adversarial testing activities to focus on 21 predefined categories (Storchan et al., 2024), while a number of major AI developers have described scoping internal pre-deployment assessment exercises around sets of defined topics and tests (often citing taxonomies like those offered by MLCommons or NIST) (Meta AI, 2024a; OpenAI, 2024).

Examples of structured taxonomies referenced in impact assessment and evaluation efforts

Human Rights (abriged)	DEFCON Red-teaming	Rights-impacting AI	MLCommons
Freedom and equality	Credit card	Civil rights, civil liberties, privacy, freedom of speech, voting rights, human autonomy, and protections from discrimination, excessive punishment, and unlawful surveillance	Violent crimes
Freedom from discrimination	AI sentience		Non-violent crimes
Right to life	Bad math		Sex-related crimes
Freedom from slavery	Citizen rights Misinformation		Child sexual exploitation
Right to seek justice	Contradictions		Indiscriminate weapons
Freedom from arbitrary arrest, detention or exile	Defamatory information	Equal opportunities, including equitable access to education, housing, insurance, credit, andemployment	(CBRNE)
Privacy and freedom from attacks on reputation	Demographic negative biases		Suicide & self-harm
Right to seek asylum from persecution	Demographic stereotypes		Hate
Freedom of thought, conscience and religion	Economic misinformation	Ability to access or apply for critical government resources or services, including healthcare, financial services, public housing, social services, transportation, and essential goods and services	
Freedom of opinion and expression	Geographic misinformation		
Freedom of peaceful assembly and association	Human rights violations		
A decent standard of living, including food, clothing, housing, medical care and social services	Known prompt injection		
Education	Legal misinformation		
	Multilingual inconsistencies		
	Overcorrection		
	Political misinformation		
	Surveillance		
	Unknown prompt injection		
	User security practices		
	Network/information security		
	AI knowledge misinformation		

Opportunities and limitations

Over a decade of work identifying potential harms of AI systems has informed the development of dozens of frameworks and taxonomies, which aim to synthesize consensus and motivate attention to the key risks, whether voluntarily or by legal or regulatory mandate. Since many calls for impact assessment and evaluation are motivated by the desire for AI developers to spot and address these sorts of harms, assessments grounded in defined frameworks can ensure that baseline categories of harm are attended to, and can add legitimacy to critique when developers fail to account for key risks (Raji et al., 2023). Structured analyses can still vary among quantitative, qualitative, or mixed methods, so multiple organizations considering the same set of topics may nevertheless differ widely across both assessment approaches and their results. Nevertheless, structuring analysis around defined taxonomies can be important to coordinating a large or disparate group of stakeholders around a shared effort (such as across multiple research or product teams within an organization, or among expert red-teamers invited to contribute to the identification of risks). It can also facilitate follow-up deliberation about which domains require further research or mitigation effort. When multiple organizations ground their efforts in similar structures, they can also encourage field-level progress by helping organizations compare their approaches and adopt methods that have proven useful in similar contexts.

These frameworks and taxonomies can themselves be consequential sites of political negotiation or consensus-building, and efforts that build on the results of these conversations can ensure the expertise and advocacy that informed their development continue to shape evaluation efforts. For instance, human rights frameworks are the result of hard-fought international conversations spanning a broad array of stakeholders and reflect an enduring consensus on a set of rights that both governments and private actors ought to protect and respect.

AI practitioners have reflected that in the absence of established taxonomies, significant time is wasted deliberating about what topics should be prioritized, how effort and resources should be allocated, and how coverage and progress should be monitored.

Structured analysis does not necessarily (and in fact, rarely) means that methods for assessment are fixed; rather, it guides what harms and impacts are considered and may set norms for methodology while still leaving some room for exploration.

Example goals that may benefit from this approach:

Inform

- Identify directional gaps in existing system safeguards or evaluation tools

Evaluate

- Evaluate sufficiency of organizational processes in facilitating risk management efforts
- Monitor effects of changes and interventions to system characteristics, outcomes, and impacts over time
- Evaluate an organization's practices against its claims

Communicate

- Enhance legibility of complex systems and their impacts by making results of assessments transparency and publicly accessible
- Facilitate deliberation and generate consensus around how impacts should be defined and prioritized, and the appropriate methods to detect and remediate them
- Establish credibility by documenting and disclosing adoption of expected practices

Change

- Enable customers to make informed decisions about system adoption and use based on assessment result

At the same time, simply structuring assessment activities around taxonomies cannot solve several important challenges. A structured but expansive list of topics or harms still requires significant capacity, resources, and expertise for an organization to thoroughly investigate each topic (Moss et al., 2021; Waem et al., 2024). The taxonomy or framework selected may still be incomplete — whether by failing to include or properly scope key considerations, aggregating them in a way that obscures important harms, improperly prioritizing among issues, or presenting both known and hypothetical topics as equally important. As a result, organizations may overlook categories of risk specific to the systems they are developing or point to reliance on existing taxonomies to unjustifiably defend decisions not to invest in more exploratory analysis. In the interest of retaining some flexibility within the assessment of impacts across defined categories, certain approaches to structured analysis may still leave room for organizations to adopt inadequate practices while claiming legitimacy from the external frameworks or methods they used. And some impacts may remain fundamentally contested such that their inclusion in structured analyses neither addresses fundamental concerns nor provides the legitimacy or utility that organizations or stakeholders hope for (Moss et al., 2021).

Considerations that can advance a more holistic approach

Structured assessment of systems creates opportunities for a variety of stakeholders to shape both the prioritization and implementation of efforts across multiple jurisdictions and institutions.

- If legal or regulatory requirements reference a structured framework or taxonomy, reflect on whether its development has incorporated input from experts and stakeholders.
- Where possible, consider establishing a process to regularly update the risks and impacts that are included.
- Provide stakeholders with sufficient time and multiple opportunities to offer feedback on new foundational frameworks.

- Explore creating channels for people to highlight if an established framework is leading key harms to be systematically overlooked, so that organizations can incorporate considerations of these additional impacts.
- Transparently explain how each impact was assessed in order to facilitate feedback on the sufficiency of these efforts.
- Employ both quantitative and qualitative methods to explore each topic that has been prioritized to ensure efforts are not too narrowly scoped.
- When providing guidance or expectations for organizations conducting structured impact assessments, consider articulating a mix of required investigation methods and recommended methods, as well as making clear that organizations should adopt methods relevant to their own context that centralized frameworks may not have anticipated.

Focused

Evaluation of a specific harm or impact or assessment against a procedural requirement.

Theory of change: As particular consequences of AI impacts become clear, robust investigation of these specific harms, grounded in mixed-method approaches, will enable their effective management — including facilitating understanding of when established approaches may be insufficient.

Sometimes referred to as targeted or directed evaluation (Weidinger et al., 2023), deliberate and **focused** efforts to thoroughly understand a particular risk or impact posed by AI systems have proven to be a necessary part of any risk management effort. Such assessments might include a suite of technical measurements, investigation of system components like training data or other design choices, thoughtful identification of impacted communities, exploration of human interaction with the system, and consideration of how impacts may play out at a societal level (Galdon Clavell,

2024). For instance, stakeholders concerned with discrimination in an AI system have encouraged developers to not only to test for specific mathematical disparities but also to understand the role of human behavior on the distribution of training data, consider how discretion in the application of a system's recommendations may introduce new biases, and take into account the broader ecosystem of institutions and confounding factors that influence how the system may shape people's and institutions' behavior over time (Reisman et al., 2018). Focused investigations of security risks might include specific evaluations of technical capability, but also consider human factors that could exacerbate risk, threat actor behavior, and areas of vulnerability outside of the system itself (such as supply chains or societal resilience). This scope of assessment may also consider an organization's practices against procedural recommendations or requirements (rather than evaluating the characteristics of a particular AI artifact), such as comparing a company's development processes against voluntary frameworks like NIST's AI RMF or organizational requirements enumerated in laws such as the EU AI Act.

Focused assessment in practice

Familiar examples of a focused approach can be found in bias assessments of AI systems, particularly those that extend beyond the investigation of specific metrics. A variety of academic and advocate-driven analyses of pretrial risk assessment tools, for example, built on initial quantitative evaluations to paint a fuller and more compelling picture of the many ways these systems perpetuate unfairness and injustice (Koepke & Robinson, 2018). Professor Virginia Eubanks' ethnography *Automating Inequality* offered a systemic investigation of the impact of predictive models on poor communities in the United States (Metcalf et al., 2021). And in the context of general purpose systems, developers have described efforts to involve scientists in probing systems to understand their particular scientific capabilities (OpenAI, 2024), with similar efforts undertaken to assess child safety risks that an AI model or AI-powered products may pose (Meta AI, 2024b; Ofcom, 2024).

Opportunities and limitations

Focused assessments that deeply examine a generally recognized issue while avoiding oversimplification are a crucial dimension of AI assessment and accountability. Experts have noted that such inquiry can provide "thick description" (or in the context of AI, inform "thick alignment") that reflects crucial contextual analysis necessary to inform meaningful understanding and intervention (Alondra Nelson, 2023; Costanza-Chock et al., 2022). Moreover, this sort of holistic inquiry provides "narrative depth" that compellingly melds quantitative and qualitative findings, helping to motivate investment or action (Vecchione et al., 2021). These sorts of efforts can also inform more specific investigations and support assessments of the validity of more precise assessment methods by illuminating relevant context and potential confounding factors that must be taken into account. And they can highlight aspects of the impact in question that have not been sufficiently recognized or prioritized through existing assessment approaches.

Example goals that may benefit from this approach:

Inform

- Inform development of reliable and valid tests to measure system behavior and impacts
- Clarify likelihood and magnitude of harms in order to prioritize allocation of resources to further research into harms and impacts and to developing effective mitigation methods

Evaluate

- Demonstrate existence or extent of issue to motivate attention to and investment in remediation
- Inform recommendations for specific changes to system design, implementation, or mitigations
- Evaluate sufficiency of organizational processes in facilitating risk management efforts
- Ensure validity of tests and benchmarks

Communicate

- Share findings with regulators and researchers to support further research on and broader understanding of systems and their impacts

Change

- Create conditions for public scrutiny for subpar practices
- Lay foundation to justify legal remedy for people who experience harmful impacts
- Stimulate public pressure toward actors to address harms that have not been sufficiently attended to
- Generate policy pressure to stimulate additional investment in harm mitigation by highlighting gaps in existing incentive structures and governance tools
- Motivate and/or justify imposition of penalties for harms not previously enumerated
- Generate evidence to establish standing for legal or regulatory challenge

Flexibility in how particular risks are explored is a desirable feature of focused assessment, as it helps to avoid framing traps (Vecchione et al., 2021), oversimplification of analysis (Casper et al., 2024), and improper operationalization of nuanced issues (Winecoff & Bogen, 2024). However, this flexibility presents some of the same vulnerabilities as broader forms of analysis. In the absence of defined standards or methods to investigate particular impacts, organizations may inadvertently or purposely overlook key facets of the harm in question (Casper et al., 2024), while still claiming (in a manner that may be difficult to contest) that they have investigated the issue and taken appropriate steps to address findings. On the other hand, overemphasizing methodological rigor can lead important questions to go unanswered because consensus methods have not yet been developed to tackle those

topics, and can lead valuable community-informed approaches to be deprioritized or excluded (Vecchione et al., 2021).

Focused assessment offers an appealing balance between exploratory efforts and specific ones, allowing known issues to be investigated from diverse perspectives.

The more narrowly scoped the topic of inquiry, the greater the temptation may be to further reduce analyses to a few predefined methods, one scalar value, or a checkbox exercise (Hutchinson et al., 2022). But inversely, a disorderly melange of analyses may be less effective in motivating accountability and recourse than holding to clearly defined metrics (regardless of the limitations that certain metrics may nevertheless present), and disagreement over the selected approaches can distract from and undermine accountability goals. And fundamentally, this sort of holistic, sociotechnical inquiry requires time, personnel, and resources to carry out, which may not be allocated at a sufficient level to conduct a robust analysis (Ofcom, 2024).

Considerations that can advance a more holistic approach

Focused assessment offers an appealing balance between exploratory efforts and specific ones, allowing known issues to be investigated from diverse perspectives. For this sort of approach to be most effectively deployed, actors should take inspiration from and integrate expert-recommended methods to ensure that technical, social, organizational, and societal impacts are considered in their efforts:

- Recognize that both technical and nontechnical skills and expertise are critical for evaluating and surfacing effective remedies for particular impacts of AI systems (Hutchinson et al., 2022). As researchers at Data & Society have noted, interdisciplinary expertise can help weave together higher level concepts with specific understanding of how impacts may manifest as harms (Metcalf et al., 2021).
- Transparently explain how the organization assessed particular impact in order to facilitate feedback on or challenge to the sufficiency of these efforts.

- Consider not only AI models or their component technical artifacts, but the sources of data, user assumptions and behavior, and organizational and societal dynamics that may influence a system's impacts. In particular, social science research methods and a focus on human interactions with a system (both how users deploy and prompt systems, and how they interpret or act on system outputs) can reveal critical dimensions of analysis that a technical artifact-focused approach will overlook (Lam et al., 2023).
- Engage with external domain experts in the topic of interest, and invest in explaining the relevant system and technical details to them in order to facilitate constructive collaboration. Such engagements are important both to design assessments and evaluations as well as to ensure analysis of results remains sound.
- Integrate context about social and historical structures of harm to inform the selection of evaluation methods and ensure analysis does not overlook systemic factors (Radiya-Dixit, 2025).
- Consider co-defining the scope of inquiry with external experts and affected communities to help ensure resulting insights comprehensively address issues most likely to matter, and co-executing research to spot methodological gaps that emerge to ensure the analysis of results captures the most relevant impacts (Vecchione et al., 2021).
- Avoid reducing analysis to overly simplistic metrics — even when it may be appealing to help prioritize attention or facilitate decisionmaking — since quantitative signals may fail to capture important context. A mixed-methods approach that incorporates both quantitative and qualitative insights will be most informative for focused analyses.
- Assess a system in its actual context of use, rather than (just) in a pre-deployment vacuum (Vecchione et al., 2021). While pre-deployment tests can provide some insight into potential impacts or harms, model-level measurements often have only tenuous relationships with downstream impacts, and are generally not able to account for contextual factors that affect how harms or impacts may manifest (Winecoff & Bogen, 2024).

Specific

Analysis of a specific harm or impact using a defined benchmark, metric, or requirement.

Theory of change: Assessing a system against defined standards can be highly compelling in leading to specific actions (e.g., motivating voluntary risk mitigations, triggering required remediation efforts, gating market access, or activating other accountability measures).

Some assessments and evaluations of AI systems involve investigating a **specific** characteristic of an AI system, such as running a predefined bias evaluation or attempting to detect whether a model exhibits a particular "capability." While such inquiries often involve quantitative analysis, such as measuring a particular metric (such as a machine learning classifier's performance) and comparing it against a benchmark or threshold (Raji et al., 2020), they can also be more qualitative — assessing whether there is sufficient evidence to back up a claim, for instance (NTIA, 2024).

Researchers have drawn comparisons between specific AI audits and *hypothesis testing* in scientific research, where researchers conduct controlled experiments to investigate a particular effect, and seek to determine whether the effects observed in an experiment are likely meaningful or simply due to random chance. Hypothesis testing is a well-established method in empirical research, and can help auditors quantify the uncertainty in their data — which is crucial for making informed decisions and developing action plans.

Learn more about the key ideas behind hypothesis testing, how it can be applied to AI audits, and the conditions where it might fall short in the companion brief to this report, Hypothesis Testing for AI Audits (Winecoff, 2025).

Specific assessment in practice

Examples of specific testing of AI systems abound, with many rooted in the legacy of systematic audit studies of housing discrimination which aimed to detect whether housing providers denied or offered different terms for housing opportunities on the basis of protected class by conducting controlled experiments in the field (Vecchione et al., 2021). Tests of disparate impact in the employment context generally involve conducting quantitative tests to determine whether selection rates fall short of an 80% ratio across protected groups (with more specific statistical significance tests generally required as part of subsequent legal challenges that may result). New York City's Local Law 144 was one of the first examples of a concrete regulation requiring independent algorithm auditing, requiring providers of automated employment decision-making tools (AEDTs) to commission measurements of their systems' "impact ratio," or the selection or scoring rate for a demographic category divided by the selection or scoring rate for the the most preferenced category (Groves, 2024; Notice of Adoption of Final Rule for Use of Automated Employment Decision-Making Tools, 2023).

Prominent efforts to test consequential AI systems, like the Gender Shades investigation of comparative accuracy of face-based gender identification systems, have similarly focused on a defined metric (in that example, error rates across gender-skin tone cohorts) despite being motivated by more systemic concerns about the harms of facial analysis tools. The researchers credit this approach with motivating three major providers of facial analysis technology to substantially reduce the performance disparities the initial study had identified (Raji & Buolamwini, 2023).

Some specific assessment efforts, meanwhile, aim to validate that an organization has adopted particular, required approaches. For instance, within EU AI conformity assessments, an evaluation may consider whether an organization that develops high-risk AI systems has adopted a specific practice that is required by the law, such as keeping sufficient technical records.

Finally, specific testing can consider whether a system or its safeguards performs as claimed. In a consequential legal challenge of AI-driven discrimination, the US DOJ required Meta to engage a third-party reviewer to verify whether the company's personalized ad delivery system was in compliance with a set of metrics outlined in the parties' negotiated settlement over claims of discriminatory housing ads (*United States v. Meta Platforms, Inc., f/k/a Facebook, Inc.*, 2022). And NIST's ARIA initiative includes an explicit goal of testing whether models submitted for testing (and any applicable safeguards) perform as claimed (NIST, n.d.).

The above examples are not necessarily models for how specific analysis should be approached, but together they illustrate notable characteristics of this sort of effort.

Opportunities and limitations

Compared to more exploratory and ad hoc approaches, a primary potential benefit of specific assessment is methodological specificity and standardized interpretation (Vecchione et al., 2021) — though importantly, adoption of a quantitative metric does not guarantee that the selected measurement method is indeed rigorous (Winecoff & Bogen, 2024). Nevertheless, specifically scoped inquiries tend to be important to clarifying assessors' tasks and making assessment efforts more legible to relevant stakeholders (which can be particularly compelling in motivating action and intervention) as well as identifying and galvanizing constructive intervention (Birhane et al., 2024). Specific assessments are more easily comparable against baselines, standards, required thresholds, and results from other systems or actors (Ofcom, 2024), though their utility can depend on whether such baselines and standards exist. A notable portion of AI accountability proposals presume the existence of techniques an assessor can deploy, or a benchmark against which they can compare a system's characteristics or an organization's practices (Raji et al., 2023), but in practice the data and methods needed to conduct these sorts of evaluations may not exist off-the-shelf. If a system has already been in operation or it is simple to apply to an input dataset, the data to conduct these measurements may be readily available (though this is not always the case even when basic information about a system's operation is available (Bogen, 2024)). In other cases, as with generative AI systems that have not yet been broadly deployed, additional efforts like red-teaming or manually constructing prompts and grading responses may be needed to generate new data to facilitate empirical testing (Storchan et al., 2024).

Presuming specific methods or inquiry are reasonably valid — that is, do they measure what they are intended to measure — they can provide concrete or quantitative signals that can be directly compared against one another or monitored over time to inform the efficacy of mitigation efforts. Even if specific assessments have methodological limitations, they can still be useful to accountability actors and enforcement agencies to motivate and guide further inquiry or investigation (Jones et al., 2024).

Example goals that may benefit from this approach:

Inform

- Benchmark against baselines and peers

Evaluate

- Trigger policies that require reduction of risk to a reasonable degree prior to further development or deployment, including implementation of auxiliary safeguards to address residual risk that primary mitigations cannot eliminate
- Assess systems against predefined thresholds in order to make decisions about mitigations or deployment
- Evaluate sufficiency of organizational processes in facilitating risk management efforts
- Ensure validity of tests and benchmarks
- Monitor effects of changes and interventions to system characteristics, outcomes, and impacts over time
- Assess whether a system has exceeded a defined threshold (e.g. precision, outcome disparity, etc)
- Evaluate an organization's practices against its claims
- Verify conformity with procedural requirements

Communicate

- Enable customers to make informed decisions about system adoption and use based on assessment results
- Demonstrate soundness of practices by disclosing favorable results and/or effectiveness of remediation

Change

- Gate market access based on verified conformity with specific metrics or tests

- Require withdrawal or decommissioning of systems that exceed defined thresholds
- Justify imposition of penalties for harms not previously enumerated
- Enable customers to readily assess potential vendors against defined procurement policies
- Generate evidence to establish standing for legal or regulatory challenge
- Provide evidence to support factual arguments for legal or regulatory challenge
- Justify imposition of penalties by demonstrating relationship between system/actor and undesirable impact

Despite the important goals that specific assessments can facilitate, approaches in this general category present a large variety of limitations that must be taken into account, and are particularly important to center as practices and requirements for AI assurance activities crystalize. Perhaps most obviously, defining the scope of inquiry too narrowly limits which issues, impacts and harms can be detected in the first place. (Metcalf et al., 2021; Moss et al., 2021; Selbst, 2021) Even if a potential harm of a system has been properly foreseen, clumsy operationalization — that is, the translation of a more ambiguous concept it into a measurable quality — can lead the subsequent analysis to be disconnected from real-world applications and people's lived experiences, or reflect an incomplete perspective of a larger issue (a particular risk for harms that are difficult to quantify). If the scope of an assessment does not accurately reflect the underlying concern, then accountability efforts that rely on it can backfire: optimizing the selected metric can distract from efforts to address root causes of an issue, and lead the metric to lose meaning (Moss et al., 2021).

In the absence of established standards for assessments, the choice of metrics can be opaque, arbitrary, gamed, and the subject of contestation (Groves, 2024; Hutchinson et al., 2022; Jones et al., 2024). Where thresholds are applied to inform a given interpretation or action (for example, $p < 0.05$ or a demographic parity metric of < 0.8), the choice of thresholds may not prove to be meaningful, which can introduce another source of artificial confidence in such approaches. On top of validity challenges, these sources of methodological instability suggest the process of arriving at a set of consensus methods will be challenging and drawn out as particular practices are debated in scientific communities and formally challenged through legal processes or enforcement actions (Hadfield & Clark, 2023).

For example, in a fair lending monitorship of lending company Upstart, the monitor and the company ultimately found themselves at an impasse over "the appropriate and legally required methodology" to determine whether a viable lending model existed that would result in less disparate impact than a baseline model but with equivalent accuracy — with the disagreement hinging on how measurements ought to incorporate statistical uncertainty (Relman Colfax, 2024).

Measurements that rely on improper sampling methods, assert definitive conclusions while failing to report on or acknowledge statistical uncertainty, or engage in *p*-hacking (manipulating tests to elicit statistically significant results) ought not be relied on, but without procedures in place to review for these issues, those being presented with the results may not be aware of the flaws in the underlying measurement (Hutchinson et al., 2022; Vecchione et al., 2021). (This can be particularly problematic in legal settings or government procurement contexts, where courts or government agencies may not have expertise to interpret nuanced quantitative evidence (Grimes, 2023).) In the context of generative AI systems, many evaluation efforts don't account for the way the model outputs can be highly sensitive to minor changes in prompts (Winecoff & Bogen, 2024), and there are generally few guarantees against contamination (i.e., instances in which test data was used as part of the model training) in evaluation and benchmarking exercises, creating concerns for measurement validity (Jones et al., 2024). And assessors may not have access to data they

need to conduct measurements in the first place, whether due to developers' hesitation to share this data or limitations in collecting it in the first place (Bogen, 2024; Groves, 2024).

Researchers have also noted that the appeal of specific assessment and the methodological rigor it implies risks undermining the importance of qualitative insights and eroding the possibility of community participation in exploring a system's impacts and harms (Jones et al., 2024; Vecchione et al., 2021). If an assessment results in binary or numerical outputs, those metrics might appear to be easily combined or compared, when in fact the act of simplification or quantification can obscure meaningful differences or lead to oversimplification of cost-benefit analyses (Hutchinson et al., 2022). In the pursuit of rigor by relying on established standards and methods, researchers can end up discounting important insights that are inherently difficult to quantify and that may be relevant to different communities, cultures, or geographies that are not sufficiently reflected in those methods (Jones et al., 2024; Solaiman et al., 2024).

Considerations that can advance a more holistic approach

Inclusive, sociotechnical, and participatory approaches to investigating systems are often referenced as important tools for more exploratory inquiry into AI systems, but they remain no less important at the more granular level of specific assessment and accountability activities:

- Specific assessment should extend beyond exclusive focus on models and technical artifacts; such efforts should also consider human interaction with those systems and inspection of organizational policies or processes against claims, standards, or expectations (Groves, 2024).
- Assessments should incorporate input from affected communities into the way that metrics are defined, selected, and interpreted to help ensure validity of those metrics to the impact and harms in question, and the applicability of those approaches to relevant societal contexts (Hutchinson et al., 2022). Researchers should directly involve affected communities in the testing and evaluation of AI systems where possible, including in field experiments (Vecchione et al., 2021). Where such practices are not adopted voluntarily, policymakers could consider incentivizing them.

- Even if a particular measurement or accountability effort is oriented towards specific metrics or endpoints, assessors should take note of relevant impacts that fall outside of that scope and ensure those gaps are made known to accountability stakeholders (Raji et al., 2020). Both quantitative and qualitative findings should be incorporated into any assessment documentation or reporting (Vecchione et al., 2021).
- To the extent standards are set that define specific methods or thresholds, those thresholds should be revisited periodically to ensure they remain relevant and incorporate recent insights as well as input from impacted communities (Raji et al., 2023). Where market entry or other privileges are gated by conformity to a quantitative metric, this metric should be frequently reviewed to ensure it is leading to sound decisions, and a broad community of experts and public stakeholders should be involved in the determination of thresholds (Groves, 2024). In some cases, appropriate thresholds may vary by context.
- Reviewers should acknowledge any assumptions made in the selection or implementation of measurement or assessment methods, including assumptions about context, data distribution, and model or system characteristics, as well as the underlying hypotheses tested (Raji et al., 2020; Vecchione et al., 2021). Reviewers should likewise acknowledge statistical limitations or uncertainty that may affect the conclusions of their inquiry, and consider providing guidance about how to interpret results in light of these constraints.
- In addition to direct community collaboration, consider opportunities to invite external stakeholders to contribute ideas or methods for model evaluations to ensure assessment efforts provide sufficient coverage for relevant harms and impacts and provide opportunities for public interest input into the assessment of evaluation results (Anthropic, 2024c; Jones et al., 2024).

***

Clearly, the scope of inquiry for a given assessment or evaluation will have a meaningful impact on what is discovered, and what actions the findings may motivate. Evaluation efforts may explore different scopes of inquiry in sequence or in parallel, but more

precisely identifying the intended scope of a particular effort or policy proposals can help stakeholders spot potential gaps and opportunities — including identifying where lower and higher degrees of independence may be particularly important.

Degree of Independence

Experts and scholars have thoroughly explored the need for robust accountability efforts to involve independent, external reviewers who can avoid conflicts of interest. Analyses of audit and impact assessment efforts often differentiate between **first party**, or a company assessing its own products or practices; **second party**, where reviewers have a contractual relationship with the auditee; and **third party**, or assessments conducted by reviewers with a higher degree of independence (Raji et al., 2022). Ultimately, though, important dimensions of independence include the degree to which an organization has control over the systems that will be included in a given inquiry and what questions may be asked about them, whether and to what extent findings are disclosed, and whether the organization conducting the inquiry has an interest in maintaining a long-term relationship with the inquiry subject or similar organizations.

Importantly, assessment efforts can vary in the degree of independence within and across these three common buckets. For instance, first party corporate governance efforts commonly incorporate what is often called the "three lines of defense": first, the developers of the product or process in question are expected to manage risk; second, responsibility, compliance, or risk teams develop frameworks, tools, and other resources to enable and oversee such efforts, and third, in-house audit teams that conduct internal reviews to evaluate the efficacy of these efforts against external expectations or requirements (Astley & Regelbrugge, n.d.; Raji et al., 2020). Researchers have also noted that the line between second party and third party efforts tends to be ambiguous (Raji et al., 2022). Given this fluidity, while recognizing the level of independence exists on a spectrum and will not always cleanly divide between categories, we will proceed in assessing the theories of change, benefits and limitations through the lens of **low**

independence, **medium independence**, and **high independence** assessment and accountability efforts.

Low independence

Theory of change: Direct and privileged access to an organization or the technical systems it builds can enable thoughtful and thorough self-assessment that helps businesses proactively map, measure, and manage risk, and generates a documentation trail that can be used for further scrutiny or oversight.

Assessment and evaluation efforts of AI systems typically begin within an organization as part of formal or informal risk management practices, ideally early on in the development lifecycle. Because organizations tend not to disclose the systems they are considering or are in the process of building before announcing their launch, such risk and impact mapping exercises necessarily start within the circle of stakeholders who have direct and privileged knowledge of a given product or system under development. In some cases, including certain public sector contexts, constraints around what data or classified information is permitted to be disclosed to external parties may require increased reliance on lower independence evaluations. Even advocates for aggressive AI accountability requirements have emphasized the foundational role of self-assessments, urging they be required by public and private entities (and potential harms mitigated) prior to launch (Reisman et al., 2018).

Organizations might choose to conduct these exercises to evaluate the impacts of a system against the organization's internal policies or external regulations, manage business risk, or proactively avoid contributing to negative impacts to people and society (Moss et al., 2021). Some such efforts may even involve external actors; for instance, research organization Model Evaluation and Threat Research (METR) has partnered with advanced AI labs to develop

and conduct model evaluations, but has clarified that these efforts should not be understood as providing meaningful oversight (Barnes, 2024).

Opportunities and limitations

Experts have noted that while internal efforts may not provide a high degree of accountability, they do have the potential to create space for internal reflection on a system's impacts (which can in some cases motivate corrective action), facilitate greater ethical sensitivity and organizational capacity for responsible AI efforts, and lead to the production of documents and findings that can be reviewed or validated by actors with higher degrees of independence (Moss et al., 2021; NTIA, 2024). Familiarity with a system can support more informed tests, which may be particularly useful in adversarial exercises like red-teaming (Galindo et al., 2024). And actors with "white-box" or unfettered access to an AI system or model may have more flexibility to conduct deeper system testing to understand causality and failure modes, and directly experiment with remedy mechanisms (Casper et al., 2024) — though low-independence actors may still underestimate or fail to foresee downstream harms and be constrained in their ability to share findings or to demand that potential mitigations be implemented compared to higher independence actors with similar system access (Moss et al., 2021).

Low-independence actors are often those who are closest to a system and who have the most direct access to technical details, candid personnel, and an understanding of the organization's plans and priorities that can inform there assessment (Raji et al., 2020), but internal actors with moderate independence (such as second and third lines of defense like internal audit teams) may also enjoy a similar degree of privileged access — albeit requiring more time for internal investigation to understand systems or processes for which they are not experts. As such, in considering the utility of assessment and accountability activities, it may be helpful to differentiate between first-party reviewers whose incentives are aligned with quickly launching products and those who are incentivized to reduce business risk even at the expense of immediate financial benefit.

To maximize confidence that an AI system operates as claimed or that the organization developing it has reasonably mitigated risks, the highest degree of independent inquiry may be warranted.

Broadly speaking, low-independence efforts can be constrained by the amount of resources made available to internal teams, incentives or disincentives for staff to work on particular topics or challenge organization practices, explicit approval from organizational leaders to pursue certain lines of inquiry or share the results of tests (both among internal colleagues or with external stakeholders), and failure to appropriately interpret the results of their own work or to address risks detected through internal assessments if they don't seem to exceed the organization's risk tolerance. Indeed, research has found that companies often fail to act on negative results revealed by internal evaluations (Jones et al., 2024). With significant autonomy to define and conduct tests of their own systems and with insufficient independent scrutiny, organizations may be prone to select a priori those assessment methods most likely to result in favorable outcomes, or highlight favorable results while understating or declining to report on unfavorable ones. And legal advisors may resist the documentation of these sorts of efforts out of concern they may be demanded later by actors conducting more independent investigation or enforcement, or try to impede demands for such information even in the face of legal and regulatory challenge (*In re: Facebook, Inc. Consumer Privacy User Profile Litigation*, 2023). All of these constraints make it clear that while low-independence efforts may be helpful in facilitating risk discovery and have the potential to bolster intrinsic motivation for change, they are unlikely on their own to contribute to durable confidence that an organization's practices meet any particular external expectations.

Moderate independence

Theory of change: Verification of system characteristics or business practices by a credible actor who is reasonably disinterested in the results of their assessment can motivate organizations to ensure their systems and processes meet expectations, generate confidence that reasonable efforts have been made to do so, and trigger and inform additional scrutiny.

Given the significant limitations for accountability of and confidence in AI systems and the organizations that develop them through low-independence efforts alone, stakeholders and policymakers have highlighted the role of AI "assurance" efforts, which generally involve some sort of external validation or verification of an organization's claims or practices related to an AI system or development process (DSIT, 2024). Such efforts are expected to be at least "operationally distinct" from the development of the system in question, and extend to efforts involving reasonably independent third party organizations in cases where there may be lower confidence in the reviewer's ability to shape the scope of review or the characterization or dissemination of results (Birhane et al., 2024). Some external actors may also help organizations become "audit-ready," for instance by sourcing and organizing relevant internal documentation or conducting preliminary analysis that a more independent organization can verify (Groves, 2024). The US AI Safety Institute recently secured formal relationships with OpenAI and Anthropic in order to facilitate pre-release model testing, though a closer review of the announcement suggests a more collaborative relationship; the details of the agreements were not disclosed due to "commercial sensitivities" (Mazmanian, 2024). While some might presume that mechanisms for government oversight would provide the highest degree of independence, others argue the existence of contractual relationships reduces the degree of independence. We consider this example "moderate" as a result

of the uncertainty around the nature of the contractual relationship, though it is possible such efforts shift further on the independence spectrum over time.

It is important to recognize that, like low-independence efforts, there may be key differences among similar-seeming moderate-independence assessments that can impact both their effectiveness and trustworthiness. For example, two instances of external parties auditing hiring algorithms illustrate the importance of spotting these subtle differences. In both cases the audits assessed whether a company's claims with regard to algorithmic bias were accurate, but they differed in the details of the auditor's relationship with the audit target. In one case, AI auditing company ORCAA was retained to audit AI hiring company HireVue's video interview-based predictive assessments; while HireVue published the resulting audit report, the company gated the results behind a legal agreement and the audit report made no declarations as to the degree of influence the company may have had on how findings were characterized (O'Neil Risk Consulting and Algorithmic Auditing, n.d.; Zuloaga, 2021). In the second, academic researchers engaged in a "cooperative audit" with AI hiring company pymetrics. This exercise was similarly constrained to reviewing a particular AI-powered employment assessment system using preset definitions, metrics, and other relevant parameters. However, in this case reviewers made the contract, audit protocol, data sharing agreement, non-compete agreement, and project budget publicly available, and structured compensation as a sponsored research grant to the researchers' university rather than a vendor relationship, which was paid before the delivery of audit results. In addition, the company was contractually barred from editing or limiting the dissemination of findings except to prevent disclosure of "proprietary information" which was not inclusive of audit findings (Wilson et al., 2021). (While findings from this effort resulted in a peer-reviewed paper published at a leading academic conference, it nevertheless resulted in critique from other scholars about apparent conflicts of interests (M. Young et al., 2022).)

Opportunities and limitations

These sort of external audits and evaluations are often characterized as second-party efforts, since they tend to involve a financial or contractual relationship between the auditor and auditee and are therefore seen to be susceptible to at least some degree of influence from the organization under review — whether directly or indirectly. For instance, even if an external organization could operate with full autonomy and editorial control over the results of their investigations, it's possible that an interest in securing future work with the audit subject (or other companies) could lead auditors to soften their findings.

On the other hand, a formal relationship between parties tends to afford evaluators with at least some access to internal documents, personnel, and expertise which may be necessary to validate certain claims, confirm the sufficiency of internal controls, or interact with products that have not yet been released to the public (NTIA, 2024). Indeed, efforts to probe systems without such access may risk overlooking key risks or misinterpreting findings. A formal relationship also provides opportunities for companies to fairly compensate stakeholders for their work even as it raises concerns about independence. As such, moderate independence efforts are likely to play an important role in the ecosystem of AI accountability — though those efforts are more likely to confer trust the more independence evaluators can demonstrate. Attempts to establish more structured professional codes of practice and accreditation processes for evaluators could help alleviate some of these concerns, but the efficacy of such professionalization will depend largely on the substance and enforceability of such standards.

It is worth noting that moderate independence assessment activities like bringing in external red-teamers (particularly domain experts) can be expensive, especially when they involve appropriate compensation (Galindo et al., 2024). Highly capitalized organizations may be in a position to direct investments toward these efforts, but smaller organizations are likely to face more challenging tradeoffs when allocating resources.

Levers that could increase independence:

- Codes of practice, professional standards, and/or accreditation for auditors and evaluators (Raji et al., 2023)
- External fora to adjudicate disagreements around organization/system access or measurement methods
- Whistleblower protections to highlight improper testing protocols or auditor-auditee relationships
- Compensation norms and standards to reduce risk that financial relationship leads to conflict of interest
- Payment for auditing or evaluation services not contingent on findings
- Reduced control by audit subject over characterization or dissemination of audit findings
- Greater latitude by reviewers who are granted privileged access to share unfavorable findings
- Healthy competitive ecosystem of auditors and evaluators so organizations don't face vendor lock-in
- Standards to ensure auditors reduce reliance on proprietary methods than cannot be independently validated (Bhatia & Allen, 2023), and increase interoperability among audit providers to reduce vendor lock-in
- Results include explanation of efforts to ensure rigor and objectivity (DSA Delegated Regulation on Audits, 2023)
- Limitations on auditing organizations from providing non-audit services to auditees (Digital Services Act, 2022)

High independence

> Theory of change: Impartial efforts to probe and validate the claims of systems and organizations — without constraint on the scope of inquiry or characterization of their findings — are necessary to surface relevant risks and to generate the necessary external pressure to ensure they are sufficiently prioritized.

To maximize confidence that an AI system operates as claimed or that the organization developing it has reasonably mitigated risks, the highest degree of independent inquiry may be warranted. High independence efforts are often conducted adversarially, sometimes without the knowledge or involvement of the entity under review, and have the fewest constraints on the scope of inquiry or the sharing of results. Actors with the necessary degree of both expertise and independence to conduct quality high-independence inquiries have tended to be journalists, academics, research institutions, and advocacy organizations, and results of their efforts have tended to be shared most openly.

Reviews of organizations or systems that result from an adversarial legal or regulatory process have also been perceived as more independent, even if reviewers have some degree of privileged access. For instance, New York City's Local Law 144 defines independent auditors as actors not involved in using or developing the automated employment decision tool in question, with no employment relationship with the AEDT vendor, and with no financial or material interest in the organization subject to the audit (Notice of Adoption of Final Rule for Use of Automated Employment Decision-Making Tools, 2023). Meanwhile, Meta's settlement with the DOJ required the company to engage a third party reviewer to verify quarterly compliance reports for conformity with a set of negotiated metrics that were codified in the settlement agreement. The company was granted leave to propose a "qualified, objective,

independent third-party professional or firm that has, at a minimum, expertise with respect to algorithmic fairness" who had not been previously engaged on the matter in question. The government could then consent to the proposed reviewer or appeal to the court to select a different one. The settlement agreement laid out the information that Meta would need to provide to the reviewer, and included a mechanism to escalate disputes to the court (*United States v. Meta Platforms, Inc., f/k/a Facebook, Inc.*, 2022).

Opportunities and limitations

High independence inquiries are likely to result in the greatest public trust in audit or evaluation outcomes, since the organization under review has the least opportunity to influence their findings. Given the relative lack of representation by communities most likely to be impacted by AI harms in in-house technical teams, high independence efforts may be better positioned to incorporate a pluralistic array of stakeholders, approaches and perspectives (Weidinger, Barnhart, et al., 2024). And because organizations developing AI technology generally have strong financial interests in launching their products to market, a lack of robust, countervailing pressure from accountability actors empowered to freely raise concrete concerns about these systems' impacts means they risk malfunctioning, harming communities, and undermining trust in the organizations building and governing them.

[J]ournalists and independent academics tend to struggle to access information they need to fully understand the systems in question.

However, external actors without formal relationships to the organization under scrutiny (including as part of legal or regulatory processes) such as journalists and independent academics tend to struggle to access information they need to fully understand the systems in question (Birhane et al., 2024; Casper et al., 2024). Their efforts tend not to have visibility into internal details about the organization or system in question and so may only be able to probe public versions of a system and review the resulting outputs, which can lead to potentially incomplete or misleading insights (Raji et al., 2020). For instance, a system may appear to satisfy statistical tests for bias, but still rely on unacceptable stereotypes to arrive at its conclusions in a way that is not apparent via output testing

(Casper et al., 2024). High independence assessments are also usually only feasible after a system or product has been released — when harms may have already manifested and remediation efforts may be more difficult — since they generally lack privileged access to unreleased technologies. As a result, these efforts provide insufficient opportunity to challenge the adoption or deployment of that system before it leads to negative impacts (Reisman et al., 2018). And if systems are updated frequently, high independence actors may not realize when a material change has taken place or have the ability to easily replicate their assessments (Ofcom, 2024). Given these limitations, experts and advocates have encouraged policymakers to push AI developers to build more channels for independent researchers to access and probe systems without undue hindrance from the system developers (Casper et al., 2024; Nicholas, 2024). Efforts that take place under the auspices of legal or regulatory processes may face fewer such challenges, but must still overcome lack of familiarity with underlying systems and in some cases, pushback from the organizations under scrutiny about what details are in scope to be made available to the independent actor.

Expectations that external reviewers eschew financial relationships with their audit targets also raise questions about how this important work will be funded; experts have encouraged regulators to consider whether public funding could support researchers' and affected communities' efforts (Reisman et al., 2018).

04 Designing Effective Assessments

As examples throughout previous sections indicate, assessment and evaluation efforts can shift up and down each axis of specificity and independence somewhat independently: a low-specificity effort can be conducted in a high-independence manner, while a highly specific inquiry can be at the lowest level of independence and still lead to useful and actionable insights. But ultimately, the ability of different efforts in driving any of the particular goal(s) described in Table 3 relates to where they sit at the intersection of these dimensions; the different shades of gray in the framework below roughly correspond to the goals such efforts may be more likely to support.

(To be sure, factors like whether assessments occur pre- or post-deployment, or whether their focus is on technical artifacts, sociotechnical systems, or organizational processes, will impact what those efforts will be able to achieve, but we leave an exploration of these additional factors aside in this analysis for the sake of clarity.)

With this framework in mind, we offer the following recommendations to practitioners and policymakers as they continue to invest in efforts to explore the impacts and risks of AI systems and to create an ecosystem of accountability to ensure this work is prioritized.

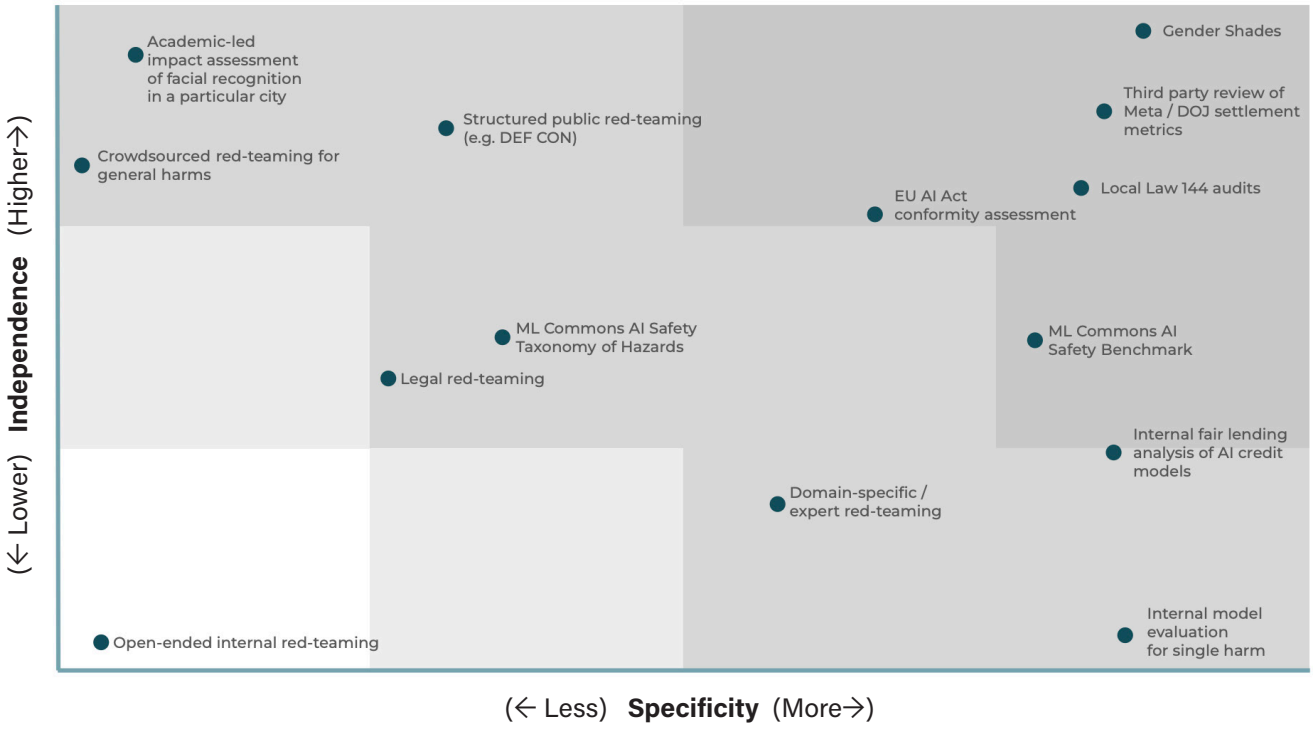

Figure 2. Mapping examples of AI assessment activities to the specificity-independence framework.

Evaluation and assessment efforts should be scoped intentionally to support a defined set of goals.

- **Practitioners and policymakers should be particularly attentive to whether the independence and/or specificity of their assessment and evaluation activities or proposals are well matched to the goals they have for those efforts**. General requirements may incentivize issue discovery, but fail to motivate behavior change or impart accountability for harms, while overly specific goals may create the conditions for stronger safeguards and oversight but reduce capacity for and interest in identifying negative impacts that narrower requirements have missed (Costanza-Chock et al., 2022). Proposed requirements for assessment and evaluations should explicitly articulate their intended goal(s) so stakeholders can assess the capacity for the proposed approach to advance that aim (and deliberate about whether that is the appropriate aim in the first place).

- **Although pre-deployment assessments are critical, the most effective assessments of AI systems will take place in the context of how those systems are implemented.** Exploratory efforts to understand these systems earlier in their development lifecycles may be important to helping developers and their institutions foresee issues when there is the greatest opportunity to mitigate them, but the most actionable insights are likely to come from more context-specific and independent inquiry after the systems have been deployed into the contexts it will actually be used. Therefore both should be prioritized.
- **Exploratory and structured inquiries should be actively leveraged to feed into more specific assessment efforts, while specific efforts should always be augmented by more exploratory inquiry.** Because efforts across the specificity spectrum offer distinct benefits and limitations, multiple approaches should be channeled into a feedback loop to surface the most comprehensive understanding of systems and motivate the most robust efforts to address negative impacts.
- **While specificity can be effective in motivating action and driving accountability, practitioners and policymakers should not pursue specificity at the expense of broader inquiry.** Exploratory efforts are important for revealing systemic and compound impacts that narrower assessments will not be able to surface (and inform the development of methods to do so systematically) so space must be preserved for more holistic approaches (Birhane et al., 2024).
- **Requirements to engage in "red-teaming" activities are often too vague to be useful, and red-teaming is not a mitigation unto itself.** Red-teaming describes the use of adversarial approaches to probe systems, but simply stating that this method has been or should be used does not reveal sufficient information about the scope or independence of an inquiry to assess whether it is likely to be effective against a particular goal. Moreover, for red-teaming to effectively support risk management, developers and their organizations must commit to (and invest sufficient resources in) fixing the issues that red-teaming exercises reveal — for instance, by retraining a model after harmful data has been removed, soliciting additional human labels to fine-tune a model, or integrating safety filters (Ofcom, 2024).

Stakeholders involved in evaluation and assessment efforts should be transparent and clear about their goals, methods, and resulting recommendations or actions.

- **Auditors and assessors should clearly disclose the methods they have employed, any assumptions that shaped their work, and what version of a system was scrutinized.** Because assessment and evaluations methods have yet to settle into a clear consensus, understanding the robustness of any given exercise against its intended goal(s) requires a candid sense of the approaches used and their limitations (NTIA, 2024). The more specific the inquiry, the more that actual or perceived methodological weaknesses may undermine confidence in the assurance these efforts are intended to confer (Anthropic, 2024b).
- **For specific evaluations, define the range of acceptable results or threshold that would pose a concern prior to conducting the test.** The value of specific tests comes largely from how the results compare against an expected value above or below which a set of actions are triggered (e.g., prompting additional tests, requiring fixes, or indicating unacceptable risk); failing to define these thresholds beforehand significantly reduces the likelihood such tests will lead to action. It also risks assessors having too much leeway to provide post-hoc rationalization for unfavorable results (Jones et al., 2024).
- **Findings from lower-independence and higher-independence efforts should flow between internal and external actors to create constructive feedback loops.** For instance, high independence exploratory or structured inquiries can surface issues that low independence activities missed, so formal channels should be established to channel findings to system developers.
- **Policymakers should advance proposals that incentivize institutions to act on and not ignore or deprioritize these findings**. Auditing and assessment are critical to identifying risks but mere identification is not the goal; those risks then need to be reasonably mitigated. Providers should not be able to rest on having merely performed assessments, without also implementing and documenting accompanying mitigations where warranted.

Accountability efforts should include as broad an array of participants and methods as feasible, with sufficient resources to ensure they are conducted robustly.

- **AI assessment and evaluation activities should include a pluralistic set of approaches that are not constrained to practitioners with technical expertise.** AI systems are not just technical artifacts, but sociotechnical systems where people, institutions, and technologies interact; understanding the impacts of these systems therefore requires interdisciplinary and mixed-method approaches (Metcalf et al., 2021). That said, certain assessments on the specific end of the spectrum may require expertise in a relevant discipline, such as statistics or law.
- **Robust audits and assessments require sufficient funding, time, personnel, and infrastructure, with compensation structures that support meaningful participatory approaches and higher independence efforts.** In the absence of clarity around expected assessment and evaluation breadth and methods, it can be tempting for organizations to revert to minimum viable efforts to demonstrate sufficient attention to external concerns. But competitive dynamics in the AI market mean that low-independence efforts are often compressed into rushed exercises by understaffed teams, and second-party audits risk being undermined by actual or perceived conflicts of interest (Casper et al., 2024; Moss et al., 2021). In such environments, participatory methods are also at higher risk of defaulting to performative activities or extractive dynamics. Practitioners should take additional care to avoid this pattern, and public funding should be channeled to independent accountability actors and technical infrastructure to support vigorous external scrutiny.

Many of the recommendations on the topics covered in this report overlap with recommendations experts have made around advancing more holistic and participatory involvement in AI. The following resources offer rich practical suggestions on this broader theme:

- CDT Blog Post: Applying Sociotechnical Approaches to AI Governance in Practice (Bogen & Winecoff, 2024)
- CDT Blog Post: Adopting More Holistic Approaches to Assess the Impacts of AI Systems (Radiya-Dixit, 2025)
- Partnership on AI Draft Guidelines for Participatory and Inclusive AI (Park, 2024)

05 Conclusion

Ultimately, no one set of accountability actors, single scope of assessment, or particular degree of auditor independence can simultaneously accomplish all of the goals that stakeholders have for AI assessment and evaluation activities (Metcalf et al., 2021). Instead, a constellation of efforts will be needed to advance this extraordinarily wide array of goals — from research, to assurance, to harm mitigation, to enforcement — that span the tangle of efforts currently described as audits, impact assessments, model evaluation, and red teaming. Without precisely parsing these goals, though, efforts that intend to advance accountability may be watered down to voluntary research exercises, while important exploratory assessment may be boxed out in favor of structured, moderate-independence exercises that neither necessarily confer the assurance stakeholders desire nor effectively surface previously unidentified consequential impacts and harms on people and society. A hammer cannot play the role of a paintbrush, and wrench serves a subtly different purpose than a screwdriver; so too will a robust ecosystem for managing AI risk require the selection of the appropriate tools for the jobs to be done.

06 References

Algorithmic Accountability Act of 2023, H.R. 5628, U.S. Congress 118th Congress (2023). https://www.congress.gov/bill/118th-congress/house-bill/5628/text [https://perma.cc/XJA6-9RER]

Alondra Nelson. (2023, July 27). *Thick Alignment* [Keynote Address]. 2023 ACM Conference on Fairness, Accountability, and Transparency. https://www.youtube.com/watch?v=Sq_XwqVTqvQ [https://perma.cc/QX7U-47P3]

Anthropic. (2024a, March 25). *Third-party testing as a key ingredient of AI policy*. https://www.anthropic.com/news/third-party-testing [https://perma.cc/B76N-LBEW]

Anthropic. (2024b, June 12). *Challenges in Red Teaming AI Systems*. https://www.anthropic.com/news/challenges-in-red-teaming-ai-systems [https://perma.cc/V6ST-BLWV]

Anthropic. (2024c, July 1). *A new initiative for developing third-party model evaluations*. https://www.anthropic.com/news/a-new-initiative-for-developing-third-party-model-evaluations [https://perma.cc/KRH2-B7QU]

Artificial Intelligence Act, 2024/1689 EU (2024). https://eur-lex.europa.eu/eli/reg/2024/1689/oj [https://perma.cc/QN39-PY6Z]

Astley, P., & Regelbrugge, A. (n.d.). *Modernizing The Three Lines of Defense Model*. Deloitte United States. Retrieved December 2, 2024, from https://www2.deloitte.com/us/en/pages/advisory/articles/modernizing-the-three-lines-of-defense-model.html [https://perma.cc/945H-CJV2]

Barnes, B. (2024, May 30). *Clarifying METR's Auditing Role* [Online post]. AI Alignment Forum. https://www.alignmentforum.org/posts/yHFhWmu3DmvXZ5Fsm/clarifying-metr-s-auditing-role [https://perma.cc/P9TR-BEC7]

Bhatia, A., & Allen, A. (2023, November 20). Auditing in the Dark: Guidance is Needed to Ensure Maximum Impact of DSA Algorithmic Audits. *Center for Democracy and Technology*. https://cdt.org/insights/auditing-in-the-dark-guidance-is-needed-to-ensure-maximum-impact-of-dsa-algorithmic-audits/ [https://perma.cc/3TRD-JG52]

Birhane, A., Steed, R., Ojewale, V., Vecchione, B., & Raji, I. D. (2024). *AI auditing: The Broken Bus on the Road to AI Accountability* (arXiv:2401.14462). arXiv. https://doi.org/10.48550/arXiv.2401.14462 [https://perma.cc/RN8V-JB8J]

Bogen, M. (2024). *Navigating Demographic Measurement for Fairness and Equity*. Center for Democracy and Technology. https://cdt.org/insights/report-navigating-demographic-measurement-for-fairness-and-equity/ [https://perma.cc/WGA8-QBPB]

Bogen, M., & Winecoff, A. (2024, May 15). *Applying Sociotechnical Approaches to AI Governance in Practice*. Center for Democracy and Technology. https://cdt.org/insights/applying-sociotechnical-approaches-to-ai-governance-in-practice/ [https://perma.cc/UT8U-AGWF]

Burt, A., & Leong, B. (2024, June 20). *A Guide to Red Teaming GenAI, Part 1*. Luminos.Law. https://www.luminos.law/blog/a-guide-to-red-teaming-genai-part-1 [https://perma.cc/J622-8GQ7]

Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M., Sharkey, L., Krishna, S., Von Hagen, M., Alberti, S., Chan, A., Sun, Q., Gerovitch, M., Bau, D., Tegmark, M., … Hadfield-Menell, D. (2024). Black-Box Access is Insufficient for Rigorous AI Audits. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2254–2272. https://doi.org/10.1145/3630106.3659037 [https://perma.cc/HLE6-CW4S]

Consumer Protections for Artificial Intelligence Act, SB24-205, Colorado General Assembly 2024 Regular Session (2024). https://leg.colorado.gov/bills/sb24-205 [https://perma.cc/K5HZ-PJVM]

Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1571–1583. https://doi.org/10.1145/3531146.3533213 [https://perma.cc/YN23-LLKB]

Delegated Regulation Supplementing Regulation (EU) 2022/2065 of the European Parliament and of the Council, by Laying down Rules on the Performance of Audits for Very Large Online Platforms and Very Large Online Search Engines, 2022/2065 Regulation (EU) (2023). https://digital-strategy.ec.europa.eu/en/library/delegated-regulation-independent-audits-under-digital-services-act [https://perma.cc/X6FQ-VFKD]

Digital Services Act, 2022/2065 EU (2022). https://eur-lex.europa.eu/eli/reg/2022/2065/oj [https://perma.cc/HV5F-37CE]

DSIT. (2024, February 12). *Introduction to AI Assurance*. GOV.UK. https://www.gov.uk/government/publications/introduction-to-ai-assurance/introduction-to-ai-assurance [https://perma.cc/9QXA-6WGV]

Galdon Clavell, G. (2024). *Checklist for AI Auditing*. European Data Protection Board. https://www.edpb.europa.eu/system/files/2024-06/ai-auditing_checklist-for-ai-auditing-scores_edpb-spe-programme_en.pdf [https://perma.cc/JKY5-PLAB]

Galindo, L., Naidoo, T., Nugteren, M., & Shah, A. (2024). *Open Loop US Program on Generative AI Risk Management: Red-Teaming and Synthetic Content*. Open Loop. https://www.usprogram.openloop.org/site/assets/files/1/openloop_us_phase1_report_and_annex.pdf [https://perma.cc/Y2NK-4PGL]

Grimes, D. R. (2023, December 8). Bad Science and Bad Statistics in the Courtroom Convict Innocent People. *Scientific American*. https://www.scientificamerican.com/article/bad-science-and-bad-statistics-in-the-courtroom-convict-innocent-people/ [https://perma.cc/8XGQ-26L4]

Groves, L. (2022). *Algorithmic Impact Assessment in Healthcare*. Ada Lovelace Institute. https://www.adalovelaceinstitute.org/project/algorithmic-impact-assessment-healthcare/ [https://perma.cc/Y82Q-N5SL]

Groves, L. (2024). *Code & Conduct: How to Create Third-Party Auditing Regimes for AI Systems*. Ada Lovelace Institute. https://www.adalovelaceinstitute.org/report/code-conduct-ai/ [https://perma.cc/8FG3-GXX2]

Hadfield, G. K., & Clark, J. (2023). *Regulatory Markets: The Future of AI Governance* (arXiv:2304.04914). arXiv. https://doi.org/10.48550/arXiv.2304.04914 [https://perma.cc/9ZPM-WLLH]

Hutchinson, B., Rostamzadeh, N., Greer, C., Heller, K., & Prabhakaran, V. (2022). Evaluation Gaps in Machine Learning Practice. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1859–1876. https://doi.org/10.1145/3531146.3533233 [https://perma.cc/YT8N-4ZQR]

In Re: Facebook, Inc. Consumer Privacy User Profile Litigation, 879, 1007-3 (United States District Court, Northern District of California February 9, 2023). https://images.law.com/contrib/content/uploads/documents/403/86462/Facebook-sanctions-order-2.9.23.pdf [https://perma.cc/AZ4E-BRMK]

Jones, E., Hardalupas, M., & Agnew, W. (2024, July 26). *Under the Radar? Examining the Evaluation of Foundation Models*. https://www.adalovelaceinstitute.org/report/under-the-radar/ [https://perma.cc/9BBB-JLLH]

Koepke, L., & Robinson, D. (2018). *Danger Ahead: Risk Assessment and the Future of Bail Reform*. Upturn. https://upturn.org/work/danger-ahead-risk-assessment-and-the-future-of-bail-reform/ [https://perma.cc/DDJ5-MZ5V]

Lam, M. S., Pandit, A., Kalicki, C. H., Gupta, R., Sahoo, P., & Metaxa, D. (2023). Sociotechnical Audits: Broadening the Algorithm Auditing Lens to Investigate Targeted Advertising. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2), 1–37. https://doi.org/10.1145/3610209 [https://perma.cc/S95C-CEBC]

Mazmanian, A. (2024, August 29). OpenAI, Anthropic to collab with NIST on AI safety testing. *NextGov*. https://www.nextgov.com/artificial-intelligence/2024/08/openai-anthropic-collab-nist-ai-safety-testing/399175/ [https://perma.cc/MX82-DQUP]

Meta AI. (2024a, April 18). *Introducing Meta Llama 3: The most capable openly available LLM to date*. Meta AI. https://ai.meta.com/blog/meta-llama-3/ [https://perma.cc/TDC8-JUWE]

Meta AI. (2024b, July 23). *Expanding our open source large language models responsibly*. Meta AI. https://ai.meta.com/blog/meta-llama-3-1-ai-responsibility/ [https://perma.cc/Z4VF-VQ8L]

Metcalf, J., Moss, E., Watkins, E. A., Singh, R., & Elish, M. C. (2021). Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 735–746. https://doi.org/10.1145/3442188.3445935 [https://perma.cc/KFE7-NEU9]

Moss, E., Watkins, E. A., Singh, R., Elish, M. C., & Metcalf, J. (2021). *Assembling Accountability: Algorithmic Impact Assessment for the Public Interest*. Data & Society. https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/ [https://perma.cc/GMW4-32NA]

Nicholas, G. (2024, August 13). *Grounding AI Policy: Towards Researcher Access to AI Usage Data*. Center for Democracy and Technology. https://cdt.org/insights/grounding-ai-policy-towards-researcher-access-to-ai-usage-data/ [https://perma.cc/46BH-6J32]

NIST. (n.d.). *Assessing Risks and Impacts of AI*. NIST. Retrieved December 3, 2024, from https://ai-challenges.nist.gov/aria [https://perma.cc/EJ2L-QQCL]

Notice of Adoption of Final Rule for Use of Automated Employment Decision-Making Tools (2023). https://rules.cityofnewyork.us/wp-content/uploads/2023/04/DCWP-NOA-for-Use-of-Automated-Employment-Decisionmaking-Tools-2.pdf [https://perma.cc/CQ8C-JA8K]

NTIA. (2024). *AI Accountability Policy Report*. US Department of Commerce. https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report [https://perma.cc/P2UR-WEHZ]

Ofcom. (2024). *Red Teaming for GenAI Harms: Revealing the Risks and Rewards for Online Safety* [Discussion Paper]. Ofcom. https://www.ofcom.org.uk/siteassets/resources/documents/consultations/discussion-papers/red-teaming/red-teaming-for-gen-ai-harms.pdf [https://perma.cc/BY6C-GWD7]

O'Neil Risk Consulting and Algorithmic Auditing. (n.d.). *Algorithmic Audit Description*. HireVue. Retrieved December 2, 2024, from https://www.hirevue.com/resources/template/orcaa-report [https://perma.cc/5DJL-B8FQ]

OpenAI. (2024, August 8). *GPT-4o System Card*. https://openai.com/index/gpt-4o-system-card/ [https://perma.cc/S532-M2BP]

Park, T. (2024, September 17). Stakeholder Engagement for Responsible AI: Introducing PAI's Guidelines for Participatory and Inclusive AI. *Partnership on AI*. https://partnershiponai.org/stakeholder-engagement-for-responsible-ai-introducing-pais-guidelines-for-participatory-and-inclusive-ai/ [https://perma.cc/7G39-XP3N]

Radiya-Dixit, E. (2025, January). Adopting More Holistic Approaches to Assess the Impacts of AI Systems. *Center for Democracy and Technology*. https://cdt.org/insights/adopting-more-holistic-approaches-to-assess-the-impacts-of-ai-systems/ [https://perma.cc/G58L-V363]

Radiya-Dixit, E., & Neff, G. (2023). A Sociotechnical Audit: Assessing Police Use of Facial Recognition. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 1334–1346. https://doi.org/10.1145/3593013.3594084 [https://perma.cc/D9PJ-G4WB]

Raji, I. D., & Buolamwini, J. (2023). Actionable Auditing Revisited: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. *Communications of the ACM*, 66(1), 101–108. https://doi.org/10.1145/3571151 [https://perma.cc/B47D-N2ZR]

Raji, I. D., Costanza-Chock, S., & Buolamwini, J. (2023). Change from the Outside: Towards Credible Third-Party Audits of AI Systems. In *Missing Links in AI Governance*. UNESCO / Mila. https://unesdoc.unesco.org/ark:/48223/pf0000384787 [https://perma.cc/SJ3X-UNCA]

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. https://doi.org/10.1145/3351095.3372873 [https://perma.cc/ZCX9-NWS6]

Raji, I. D., Xu, P., Honigsberg, C., & Ho, D. (2022). Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 557–571. https://doi.org/10.1145/3514094.3534181 [https://perma.cc/C2B9-JR6W]

Reisman, D., Schultz, J., Crawford, K., & Whitaker, M. (2018). *Algorithmic Impact Assessments Report: A Practical Framework for Public Agency Accountability*. AI Now Institute. https://ainowinstitute.org/publication/algorithmic-impact-assessments-report-2 [https://perma.cc/CDK8-DR6F]

Relman Colfax. (2024). *Fourth and Final Report of the Independent Monitor: Fair Lending Monitorship of Upstart Network's Lending Model.* Relman Colfax PLLC. https://www.relmanlaw.com/assets/htmldocuments/Upstart%20Final%20Report.pdf [https://perma.cc/P5R2-JW36]

Safe and Secure Innovation for Frontier Artificial Intelligence Models Act, SB 1047, California Legislature 2023-2024 Regular Session (2024). https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB1047 [https://perma.cc/B76P-82GH]

Selbst, A. D. (2021). An Institutional View of Algorithmic Impact Assessments. *Harvard Journal of Law & Technology*, 35(1), 119–191. https://jolt.law.harvard.edu/assets/articlePDFs/v35/Selbst-An-Institutional-View-of-Algorithmic-Impact-Assessments.pdf [https://perma.cc/GN8J-3TY7]

Sendak, M., Elish, M. C., Gao, M., Futoma, J., Ratliff, W., Nichols, M., Bedoya, A., Balu, S., & O'Brien, C. (2020). "The human body is a black box": Supporting clinical decision-making with deep learning. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 99–109. https://doi.org/10.1145/3351095.3372827 [https://perma.cc/P8BA-SB5P]

Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S. L., Chen, C., III, H. D., Dodge, J., Duan, I., Evans, E., Friedrich, F., Ghosh, A., Gohar, U., Hooker, S., Jernite, Y., Kalluri, R., Lusoli, A., Leidinger, A., … Subramonian, A. (2024). *Evaluating the Social Impact of Generative AI Systems in Systems and Society* (arXiv:2306.05949). arXiv. https://doi.org/10.48550/arXiv.2306.05949 [https://perma.cc/HC6Y-YKHS]

Storchan, V., Kumar, R., Chowdhury, R., Goldfarb-Tarrant, S., & Cattell, S. (2024). *Generative AI Red Teaming Challenge: Transparency Report.* Humane Intelligence. https://www.humane-intelligence.org/grt [https://perma.cc/C6VY-FX3G]

UK Information Commissioner's Office. (2022). *A Guide to ICO Audit: Artificial Intelligence (AI) Audits*. https://ico.org.uk/media/for-organisations/documents/4022651/a-guide-to-ai-audits.pdf [https://perma.cc/HE69-RZ74]

United States v. Meta Platforms, Inc., f/k/a Facebook, Inc., 22 Civ. 5187 (Southern District of New York June 21, 2022). https://www.justice.gov/crt/case/united-states-v-meta-platforms-inc-fka-facebook-inc-sdny [https://perma.cc/F9EL-DXV8]

Validation and Evaluation for Trustworthy (VET) Artificial Intelligence Act, S.4769, U.S. Congress 118th Congress (2024). https://www.congress.gov/bill/118th-congress/senate-bill/4769/text [https://perma.cc/4X6Q-VGDH]

Vecchione, B., Levy, K., & Barocas, S. (2021). Algorithmic Auditing and Social Justice: Lessons from the History of Audit Studies. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9. https://doi.org/10.1145/3465416.3483294 [https://perma.cc/3YXS-ARGG]

Waem, H., Dauzier, J., & Demircan, M. (2024, March 7). *Fundamental Rights Impact Assessments under the EU AI Act: Who, what and how?* Technology's Legal Edge. https://www.technologyslegaledge.com/2024/03/fundamental-rights-impact-assessments-under-the-eu-ai-act-who-what-and-how/ [https://perma.cc/TE4G-XGK7]

Weidinger, L., Barnhart, J., Brennan, J., Butterfield, C., Young, S., Hawkins, W., Hendricks, L. A., Comanescu, R., Chang, O., Rodriguez, M., Beroshi, J., Bloxwich, D., Proleev, L., Chen, J., Farquhar, S., Ho, L., Gabriel, I., Dafoe, A., & Isaac, W. (2024). *Holistic Safety and Responsibility Evaluations of Advanced AI Models* (arXiv:2404.14068). arXiv. https://doi.org/10.48550/arXiv.2404.14068 [https://perma.cc/P4J8-54DC]

Weidinger, L., Mellor, J., Pegueroles, B. G., Marchal, N., Kumar, R., Lum, K., Akbulut, C., Diaz, M., Bergman, S., Rodriguez, M., Rieser, V., & Isaac, W. (2024). *STAR: SocioTechnical Approach to Red Teaming Language Models* (arXiv:2406.11757). arXiv. https://doi.org/10.48550/arXiv.2406.11757 [https://perma.cc/WD45-5V92]

Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., Gabriel, I., Rieser, V., & Isaac, W. (2023). *Sociotechnical Safety Evaluation of Generative AI Systems* (arXiv:2310.11986). arXiv. https://doi.org/10.48550/arXiv.2310.11986 [https://perma.cc/CGB5-UUW4]

Wilson, C., Ghosh, A., Jiang, S., Mislove, A., Baker, L., Szary, J., Trindel, K., & Polli, F. (2021). Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 666–677. https://doi.org/10.1145/3442188.3445928 [https://perma.cc/FV3V-BQGP]

Winecoff, A. (2025, January). Hypothesis Testing for AI Audits. *Center for Democracy and Technology.* https://cdt.org/insights/hypothesis-testing-for-ai-audits/ [https://perma.cc/B3F6-6ZQD]

Winecoff, A., & Bogen, M. (2024, March 6). Trustworthy AI Needs Trustworthy Measurements. *Center for Democracy and Technology*. https://cdt.org/insights/trustworthy-ai-needs-trustworthy-measurements/ [https://perma.cc/66KP-445Y]

Young, M., Katell, M., & Krafft, P. M. (2022). Confronting Power and Corporate Capture at the FAccT Conference. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1375–1386. https://doi.org/10.1145/3531146.3533194 [https://perma.cc/S2J7-9D7F]

Young, S. D. (2024). *Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence*. Office of Management and Budget. https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf [https://perma.cc/BN9M-VVAW]

Zuloaga, L. (2021, January 12). *Industry Leadership: New Audit Results and Decision on Visual Analysis.* Hirevue.Com. https://www.hirevue.com/blog/hiring/industry-leadership-new-audit-results-and-decision-on-visual-analysis [https://perma.cc/KAE5-5WFT]

cdt.org

cdt.org/contact

Center for Democracy & Technology
1401 K Street NW, Suite 200
Washington, D.C. 20005

202-637-9800

@CenDemTech

CENTER FOR
DEMOCRACY
& TECHNOLOGY