From Protocols to Evidence: Bounded Claims for AI in Service of the Common Good

Nitesh V. Chawla
University of Notre Dame
Notre Dame, USA
nchawla@nd.edu

Paolo Benanti
LUISS Guido Carli University
Rome, Italy
pbenanti@luiss.it

Abstract

Claims that artificial intelligence (AI) systems improve decisions, broaden access, reduce harm, or empower users can exceed what their evaluation supports. This gap motivates our ongoing work on RISE AI. In this paper, we examine the proposed framework in light of Pope Leo XIV's *Magnifica Humanitas*, particularly its treatment of human dignity, ownership, and technological power.

We develop three linked propositions. First, AI adoption can reveal unmet human needs and institutional weaknesses, while deployment can change how institutions respond to them. We propose a rupture test to examine whether deployment repairs or compounds existing failures, substitutes for human support, or conceals problems behind improved performance. The test would compare deployment with existing practice and feasible human or non-AI alternatives. Second, responsible AI must examine the institutional conditions surrounding a system, including who owns the infrastructure, controls its use, and bears responsibility for its consequences. Third, favorable evaluation cannot by itself establish respect for human dignity or moral and political legitimacy. We distinguish evidence-bounded deployment, which limits claims to what the evidence supports under specified conditions, from measurement-bounded governance, which identifies constraints that favorable results cannot override. We outline a proposed RISE AI framework for applying these propositions across Responsibility, Inclusivity, Safety, and Empowerment.

CCS Concepts

• **Social and professional topics** → **Computing / technology policy**; • **Computing methodologies** → *Artificial intelligence*.

Keywords

responsible AI, AI governance, AI evaluation, construct validity, political economy, human dignity, RISE AI

ACM Reference Format:
Nitesh V. Chawla and Paolo Benanti. 2026. From Protocols to Evidence: Bounded Claims for AI in Service of the Common Good. In *ACM AI Summit 2026 (AI Summit '26), August 31-September 02, 2026, Atlanta, GA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3806096.3844885



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
AI Summit '26, Atlanta, GA, USA
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2638-5/2026/08
https://doi.org/10.1145/3806096.3844885

1 Introduction

As institutions and individuals delegate tasks to AI, they must also judge the efficacy of those AI outputs [20], and retain the ability and authority to intervene [21]. A system may perform its assigned task well while weakening access to human expertise, making decisions harder to challenge, or obscuring who answers for harm [11, 12]. Evidence collected for one population or system version may not support the same claim in another. Evaluation must examine these effects and limits alongside technical performance. It must also consider the conditions that preceded adoption of AI, as consequences of AI also take shape through the people and institutions that develop, deploy, and use it.

These conditions matter when people turn to AI for guidance and support. AI's availability and responsiveness raise questions about what people are seeking and what their institutions or relationships may have failed to provide. If a technology that reached its current scale only recently can so quickly unsettle our understanding of what it means to be human, then the problem may not be the technology alone. In this sense, AI can be both revelation and intervention. Its use may expose failures of care, belonging, and accountability; its deployment can change how institutions respond to them. An institution may use AI to address a failure, deepen it, replace human support, or conceal an unresolved problem behind better performance measures. To judge whether AI improves a service, we need to understand where that service was already failing and whether deployment addresses those failures. Any new harms introduced by AI still require scrutiny, even when the previous service was inadequate.

AI ethics has articulated commitments to fairness, accountability, safety, and human oversight [2, 3, 8, 9]. These commitments increasingly appear in standards, legal requirements, and assurance practices [5–7, 14]. Their interpretation depends on context, and contemporary technologies and institutions may still lack adequate mechanisms for translating them into system requirements, organizational responsibilities, and evaluation practices. Where such mechanisms exist, institutions must establish whether they fulfill those commitments and what evidence supports that judgment.

Pope Leo XIV's Magnifica Humanitas informs our treatment of dignity, technological power, and the common good. It directs attention beyond the artifact to the institutional, economic, and political order in which AI is developed and used [1]. Drawing on this frame, we examine what constitutes improvement and which uses remain unacceptable regardless of measured performance.

We develop three linked propositions. First, AI adoption can reveal unmet needs, while deployment can change how institutions respond to them. We propose a rupture test to assess those changes against existing practice and feasible alternatives. Second,

responsible AI must examine the ownership and decision authority surrounding a system, including who bears the consequences of its use. Third, evidence of improved performance cannot by itself justify a deployment or establish respect for human dignity.

We propose RISE AI as a framework for applying these propositions to claims about responsibility, inclusivity, safety, and empowerment. Institutions would identify the evidence supporting each claim, who is responsible for acting on it, and the constraints that favorable results cannot override.

2 AI, Human Dignity, and the Social Order

Pope Leo XIV's *Magnifica Humanitas* frames AI as one of the *res novae* of our time, in continuity with the social transformations addressed by *Rerum Novarum* in 1891 [1, 4]. The encyclical extends the universal destination of goods to patents, algorithms, platforms, infrastructure, and data (para. 67), observes that control over platforms, infrastructure, data, and compute often lies with major economic actors that set conditions of access and participation (para. 95), and warns that AI can amplify the power of those who already possess resources, expertise, data, and regulatory influence (paras. 106–110) [1]. Technological innovations, including AI, are not neutral as they can foster participation and justice or intensify inequality, control, and exclusion (para. 85). Systems reflect the assumptions of those who design and train them (paras. 104, 111), while responsibility extends from developers and deployers to institutions, regulators, and users (paras. 105–111, 170, 209). Dignity is not one value to be optimized alongside others. It is an inalienable premise governing the legitimacy of institutions and the treatment of every person.

The encyclical also moves from these broader concerns to concrete action. Paragraph 14 connects dignity and the common good to responsible planning, human and social impact assessment, inclusion of vulnerable populations, and digital literacy. Paragraph 156 calls for verifiable measures protecting employment, retraining, and worker participation when AI is introduced. Paragraph 164 calls for consequential algorithmic decisions to be understandable, contestable, and subject to oversight [1]. These passages raise practical questions for system design: can people understand and challenge a decision, and can responsibility for it be traced? Protocols can help establish whether these protections work in practice. Judgments about whether a particular use of AI is acceptable still require moral and political reasoning.

3 From Commitments to Evidence

Commitments to human agency and accountability have practical implications for how institutions design and use AI. Laws, standards, and assurance practices already give these commitments form through requirements for oversight, documentation, and review. The following chain traces how a commitment can inform system design and evaluation, ending with a claim whose scope is supported by the evidence.

> **Commitment → Design Objective → System Requirement → Implementation Mechanism → Evaluation Protocol → Bounded Evidence Claim**.

The chain begins with a commitment the institution seeks to uphold. A design objective states what that commitment requires in the setting, and system requirements specify what must be built or put in place. Implementation mechanisms give those requirements practical form. An evaluation protocol then specifies how their operation and effects will be assessed. The findings determine which claims the institution can support and under what conditions.

Consider an institution using AI to recommend whether someone qualifies for a service. If the commitment is to preserve human agency, a design objective is to enable affected people to understand and challenge adverse recommendations. This requires an explanation they can use and access to a reviewer with authority to change the decision, without penalty for seeking review. The institution would need to record the basis of the recommendation, route challenges to that reviewer, and prevent irreversible action while review is pending. Evaluation would examine whether people understand the explanation, can complete the appeal, and obtain correction when warranted. It would also investigate who abandons the process or cannot access it. If the findings support a claim of meaningful control, that claim must remain limited to the decisions, populations, and conditions actually studied.

This translation changes the unit of responsible AI from the model to the sociotechnical system (institution) itself. An accurate model can participate in an irresponsible system if no actor answers for failure. A model that appears safe under benchmark evaluation can become unsafe through workflow integration, user overreliance, absent recourse, or weak institutional capacity. A usable interface can still disempower users who cannot understand, contest, override, or refuse its recommendations.

Design Patterns. Existing pattern catalogues describe ways to implement governance objectives in AI systems [13]. We draw on this work to discuss mechanisms for meeting those objectives and the evidence needed to assess whether they work in practice.

1) Answerability-by-Design. Consequential systems should specify who answers for which decision at each lifecycle stage through role definitions, provenance, audit logs, escalation, and redress. The rupture test asks whether AI restores institutional answerability or further diffuses it across vendors, deployers, and users.

2) Contestability-by-Design. Affected people need understandable grounds for challenge, usable appeal or refusal pathways, timely human review with authority, and protection against retaliation. Evaluation must test practical use, not merely the presence of an appeal button.

3) Agency-Preserving Interfaces. Interfaces should expand the capacity to understand options and act on self-determined goals. Useful friction and cognitive forcing functions can reduce overreliance on AI, while meaningful human control requires that people retain the ability and authority to act on their responsibilities [20, 21].

4) Context-Aware Evaluation. Benchmark results do not constitute universal evidence. Evaluation must match the deployment population, workflow, institution, and version. A general score may support a narrow capability claim without supporting a claim about its effects in a particular setting.

5) Evidence-Bounded Deployment. Deployment claims should not exceed what the evidence supports. Evidence gaps, conflicts, and expiry conditions should be recorded explicitly. They should trigger qualification of the claim, additional evaluation, remediation, or withdrawal when necessary.

EU AI Act Requirements and Mechanisms. The EU AI Act translates governance commitments into requirements for AI development and use. For high-risk systems, Articles 9–15 establish technical and organizational requirements, including documentation, logging, and human oversight. Articles 27 and 43 address fundamental rights impact assessments for specified deployers and conformity assessment, while Articles 72–73 cover post-market monitoring and serious incident reporting [7].

Evaluation must examine how these requirements work in practice. Article 26 requires oversight personnel to have the necessary competence, training, authority, and support; evaluation must establish whether they can exercise that authority in their working conditions [7]. A log may show that an appeal was processed without showing whether the person obtained meaningful review or correction. Compliance records can support these inquiries, but their existence alone does not establish that a deployment empowers people or addresses an institutional failure.

Our proposal for RISE AI focuses on what claims those records support, under what conditions, and where that support ends.

4 Limits of Design and Measurement

Evaluating the sociotechnical system brings institutional practices into view, but the scope of evaluation still matters. An evaluation methodology may establish that a deployed system preserves meaningful control for its users, offers practicable contestation, and communicates uncertainty well. At the same time, it may say nothing about who owns the compute on which it runs, who set the agenda under which it was built, or whose labor and resources sustain it [10].

Magnifica Humanitas makes these questions part of the object of governance. It treats algorithms, platforms, infrastructure, and data as goods whose concentration can violate their universal destination; identifies private control over data and compute as a source of political and economic asymmetry; and asks who can train and govern systems and who is merely subjected to them (paras. 67, 95, 106–110) [1]. Evidence supporting a claim about a deployment does not by itself establish that the surrounding distribution of power is just. Ownership, labor, resource use, and regulatory influence therefore remain subjects of governance.

Construct validity concerns whether observed indicators and empirical relationships support an intended interpretation of a construct [17–19]. We treat inherent human dignity as a moral premise. Observable conditions may provide evidence that dignity has been respected or violated, but they do not measure how much dignity a person possesses or determine whether its violation is justified by other benefits. A system can perform well on every selected measure and still treat a person merely as a means.

Conditions relevant to agency, recourse, and access can change when AI is introduced. People may gain access to a service while losing the ability to challenge its decisions or obtain human assistance. Within RISE AI, such claims would be assessed against evidence from the population and setting studied. The validity question is whether the evidence supports the claim being made; a favorable result does not establish that the deployment respects human dignity.

Table 1: Evidence-bounded claims and measurement-bounded constraints.

	Evidence-bounded claims	Measurement-bounded constraints
Question	What does evidence support here?	What may not be authorized?
Object	Variable conditions: agency, recourse, access, safety, answerability.	Dignity-based constraints, legitimate refusal, prohibited uses, non-substitutable relationships.
Evidence	Supports, qualifies, or contradicts a claim.	May reveal a breach; cannot override the boundary.
Basis / authority	Predeclared protocol and accountable decision process.	Law, policy, affected communities, institutional mission, moral judgment.
Decision	Support, qualify, remediate, retest, or withdraw.	Prohibit, stop, redesign, or require a human alternative.
Proposed RISE role	Evaluates and records evidentiary support.	Records the boundary, basis, authority, and revision process.

Evidence-bounded deployment requires institutions to limit their claims about an AI system to what the evidence supports under the conditions evaluated. We use *measurement-bounded governance* to describe a further limit: favorable results cannot justify every use of a system. Some uses remain unacceptable even when performance, such as accuracy, is strong. An institution may, for example, require access to human review regardless of gains in processing speed. An evidence profile should therefore record these constraints alongside the evaluation results, making clear which commitments those results cannot override.

These commitments set constraints on deployment. They are not constructs to be measured, although evidence may show whether they have been breached. For each constraint, the governance record should state its basis, whom it protects, and who has authority to interpret or revise it. That basis may be law, institutional policy, or a commitment made to an affected community. Article 5 of the EU AI Act provides an example: where a practice is prohibited, favorable performance does not make it permissible [7]. Table 1 summarizes the distinction between claims assessed through evidence and constraints that the evaluation must respect.

Responsible AI requires both better design and continued moral and political judgment. Moral and political judgment remains necessary because AI-mediated systems can redistribute agency and responsibility, raising questions of authorship, accountability, and authority that design alone cannot settle.

5 RISE AI as an Evidence Architecture

The proposed RISE AI framework would organize bounded system claims around the four constructs of Responsibility, Inclusivity, Safety, and Empowerment, answering four questions: Who answers for the system? Whose perspectives shape it and whose needs does it serve? Whom does it protect from harm? Who gains or loses agency through it? We propose the following validity chain for assessing each claim.

Construct → Bounded System Claim → Indicator → Evidence Source → Bounded Inference.

A claim that AI improves a service or gives people greater control requires a clear comparison with existing practice and feasible alternatives. Within RISE AI, the question would be whether the evidence supports that claim under the conditions studied. A metric is useful only if there is a defensible connection between what it measures and what the institution claims [17–19].

Our proposal builds on existing governance and assurance work. NIST AI RMF, ISO/IEC 42001, and the EU AI Act provide guidance, requirements, and obligations that institutions must interpret for their settings [5–7]. Pattern catalogues describe practices for implementing governance objectives [13], while assurance cases connect claims to arguments and evidence [14]. Documentation, provenance records, and operational logs can contribute to that evidence [14–16]. RISE AI would use these resources to examine what an institution can reasonably claim about a deployment and when changes to the system or its use require that claim to be reviewed.

We propose a versioned claim-evidence graph to connect claims with their supporting evidence. An evidence profile would present the relevant part of that graph for a particular system, context, and evaluation period. The graph would link each bounded claim to evidence artifacts, scope conditions, accountable actors, and constraints, recording evaluators' judgments about whether the evidence supports, qualifies, or contradicts the claim. At minimum, the record should identify the system, model, data, and interface version; the relevant population and context; the accountable owner; the indicator and threshold; the provenance and limitations of the evidence; and any expiry or change trigger. For each claim, the proposed record would distinguish whether the evidence is direct or proxy-based, whether it supports or conflicts with the claim, and whether required evidence is missing or stale. A *context-and-power record* would document ownership, financing, and decision authority, along with the labor and resources the system depends on and who bears its risks. This would allow reviewers to examine whether those arrangements limit the institution's ability to change or withdraw the system.

Operational Protocol and Institutional Baseline. We propose an assessment process that institutions could use alongside their existing evaluation and review procedures. The team would first define the system, the people affected, and the problem the deployment is intended to address. It would document existing institutional conditions and identify feasible alternatives, including human and non-AI approaches.

The team would then formulate claims with affected communities, domain experts, and system owners, documenting disagreements about what should count as improvement. Before evaluation, it should specify the evidence needed to assess each claim, the criteria for judging the findings, and the constraints on deployment. Evidence could come from system tests, user studies, or records of decisions and appeals, depending on the claim. The assessment would distinguish direct observations from proxy measures and report conflicting findings, missing evidence, and results that are no longer current. Changes to the system or its institutional setting would prompt review of the affected claims and the deployment or procurement decisions based on them.

What counts as timely review or sufficient understanding depends on the setting. Under the proposed framework, institutions would explain their thresholds, identify who approved them, and provide a way to challenge those choices. A claim could take the following form: *For this system version, population, setting, and period, the evidence supports [claim] under [conditions]. It leaves [questions] unresolved. Changes to [specified conditions] would require the claim to be reviewed.*

The proposed rupture test would examine what improves and what may be lost when AI is introduced. Faster processing, for example, could reduce delays while cuts in caseworker availability make it harder for applicants to obtain help. Comparison with existing practice and feasible alternatives would need to account for both effects. An institution claiming to have addressed a service failure would therefore need evidence beyond processing speed, including whether people can still obtain assistance and challenge a decision.

Public-Benefits Scenario. Suppose an AI system supports public-benefits eligibility decisions in an institution already marked by slow processing, limited caseworker capacity, and difficult appeals. Faster decisions would not, by themselves, show that applicants have better access to assistance or appeals. Table 2 shows how one RISE chain would apply to an empowerment claim based on meaningful contestability of an adverse recommendation.

Table 2: A compact RISE evidence contract for contestability.

Element	Public-benefits instantiation
Claim	Applicants can meaningfully contest adverse AI-mediated recommendations.
Requirements	AI disclosure; understandable reasons; accessible online and offline appeal; authorized human review; non-retaliation; correction of downstream effects.
Mechanisms	Decision notice; reason codes; appeal endpoint; review queue; adverse-action pause; correction and recovery workflow.
Evidence	Versioned decision, notice, appeal, review, incident, complaint, and recovery logs; accessibility and comprehension studies.
Indicators	Notice comprehension; appeal initiation and completion; abandonment; review latency; reversal; subgroup disparities; recovery completeness.
Boundary	Specified system version, jurisdiction, channels, languages, populations, decisions, and evaluation period.
Context & power	System and data owner; vendor and cloud dependencies; procurement terms; eligibility-policy authority; caseworker staffing effects; control of appeal records.
Constraints	No irreversible adverse action without authorized human review; no online-only appeal; no retaliation for contesting a decision; preserve practicable access to a caseworker.

Example Use Cases. Education and clinical decision support offer settings in which the proposed framework could be studied. In education, a claim that AI empowers students would require evidence beyond task completion or correct answers. Can students apply what they have learned without the tool? Do they retain control over their work and access to a teacher when needed? In clinical decision support, evaluation would need to examine how clinicians use a recommendation, whether they can override it, and what happens when it is wrong. Predictive accuracy alone would not establish safety in practice. In either setting, any resulting claim would need to specify whom the evidence covers and the conditions under which it was collected. These examples identify questions for future studies; they do not establish that RISE AI is effective in either domain.

6 Research Roadmap

The next phase requires detailed operationalization, implementation, and empirical testing of RISE AI. Seven questions matter most. *Participatory content validity* asks whether affected communities, domain experts, and accountable institutions judge bounded claims to cover what matters and where their definitions diverge. *Inter-evaluator reliability* tests whether independent evaluators classify the same evidence, gaps, and inference boundaries similarly. *Diagnostic value* examines whether RISE AI profiles reveal unsupported claims, evidence gaps, or failures missed by benchmarks and compliance review. *Change sensitivity* asks whether claims are appropriately revised or invalidated when underlying data, models, deployment or use contexts change. *Decision utility* asks whether evidence profiles improve the justification of procurement, deployment, and remediation decisions. *Political-economic sensitivity* examines whether the context-and-power record exposes dependencies or distributions of risk and benefit that system-level assessment misses and whether those disclosures affect decisions. *Operational burden and proportionality* examines the expertise, time, infrastructure, and cost required, including whether credible implementation remains feasible beyond large organizations.

Initial studies could use draft evidence profiles and manual review; testing the full process would require an implementation of RISE AI, which is part of future work. This implementation should also sit alongside existing requirements repositories, audit systems, observability tools, and governance workflows rather than replace them. Building it will test whether the pieces of the framework actually work together, what burden they impose, and where the framework itself needs to change.

Limitations. The examples illustrate the proposal; they do not validate RISE AI. Claims, baselines, and thresholds depend on who participates and who has decision authority. An inadequate baseline can make a weak deployment appear beneficial. Evidence may underrepresent those most affected, and its collection can create privacy risks. Recording ownership and authority also does not ensure that either will be meaningfully scrutinized.

7 Conclusion

Evaluation must examine how an AI system performs and what changes for people when an institution adopts it. Faster or more accurate decisions may still leave people with less access to human expertise or fewer ways to challenge an outcome. Some problems may predate adoption; others may result from the deployment itself. We propose a rupture test that would compare deployment with existing practice and feasible human or non-AI alternatives, helping examine whether the institution addresses the underlying problem and what harms or losses accompany any gains.

The proposed RISE AI framework brings these concerns into the assessment of institutional claims about AI. It identifies three boundaries: the comparison against which improvement is claimed, the scope of the supporting evidence, and the constraints that favorable results cannot override. The rupture test would address the comparison; validity chains and evidence-bounded deployment would define the scope of support; and measurement-bounded governance would make the constraints explicit.

Our reading of *Magnifica Humanitas* connects these questions to dignity, technological power, and the common good. The engineering and moral-political agendas must proceed together. Whether AI serves the common good depends on what people build and what institutions choose to protect. Those choices carry a responsibility to build systems and institutions worthy of humanity.

References

[1] Pope Leo XIV. 2026. *Magnifica Humanitas: On Safeguarding the Human Person in the Time of Artificial Intelligence.* Vatican.

[2] Malak Sadek, Emma Kallina, Thomas Bohné, Céline Mougenot, Rafael A. Calvo, and Stephen Cave. 2025. Challenges of responsible AI in practice: Scoping review and recommended actions. *AI & Society* 40, 199–215.

[3] Tita Alissa Bach, Magnhild Kaarstad, Elizabeth Solberg, and Aleksandar Babic. 2025. Insights into suggested Responsible AI (RAI) practices in real-world settings: A systematic literature review. *AI and Ethics* 5, 3185–3232.

[4] Pope Leo XIII. 1891. *Rerum Novarum: On Capital and Labor.* Vatican.

[5] Elham Tabassi. 2023 Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. *National Institute of Standards and Technology.* 10.6028/NIST.AI.100-1.

[6] International Organization for Standardization. 2023. *ISO/IEC 42001:2023 Information Technology — Artificial Intelligence — Management System.* ISO.

[7] European Union. 2024. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. *Official Journal of the European Union.*

[8] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9, 389–399.

[9] Lorenn P. Ruster and Jenny L. Davis. 2025. The gaps that never were: Reconsidering responsible AI's principle-practice problem. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 350–360.

[10] Kate Crawford. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.* Yale University Press.

[11] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68.

[12] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 33–44.

[13] Qinghua Lu, Liming Zhu, Xiwei Xu, Jon Whittle, Didar Zowghi, and Aurelie Jacquet. 2024. Responsible AI Pattern Catalogue: A collection of best practices for AI governance and engineering. *ACM Computing Surveys* 56, 7, Article 173, 1–35.

[14] Alpay Sabuncuoglu, Christopher Burr, and Carsten Maple. 2025. Justified evidence collection for argument-based AI fairness assurance. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 18–28.

[15] Karl Werder, Balasubramaniam Ramesh, and Sophia Rongen Zhang. 2022. Establishing data provenance for responsible artificial intelligence systems. *ACM Transactions on Management Information Systems* 13, 2, Article 22, 1–23.

[16] Patrick Loic Foalem, Léuson M.P. da Silva, Foutse Khomh, Heng Li, and Ettore Merlo. 2025. Logging requirement for continuous auditing of responsible machine learning-based applications. *Empirical Software Engineering* 30, Article 97.

[17] Lee J. Cronbach and Paul E. Meehl. 1955. Construct validity in psychological tests. *Psychological Bulletin* 52, 4, 281–302.

[18] Samuel Messick. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist* 50, 9, 741–749.

[19] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 375–385.

[20] Zana Buçinca, Maja B. Malaya, and Krzysztof Z. Gajos. 2021. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1, Article 188.

[21] Luciano Cavalcante Siebert, Maria Luce Lupetti, Evgeni Aizenberg, Niek Beckers, Arkady Zgonnikov, Herman Veluwenkamp, David Abbink, Elisa Giaccardi, Geert-Jan Houben, Catholijn M. Jonker, Jeroen van den Hoven, Deborah Forster, and Reginald L. Lagendijk. 2023. Meaningful human control: Actionable properties for AI system development. *AI and Ethics* 3, 1, 241–255.