TRUTHIFY: Building Warranted Trust in Generative AI Through Verifiability, Contestability and Institutional Accountability

Abstract

The central problem in the governance of Artificial Intelligence (AI) lies less with a lack of an ethical code for AI. Rather, it is harder to develop those codes into actionable processes that could engender warranted trust in AI. More than just formal commitment is required to do so. Evidence, process clarity, oversight, and the existence of institutions capable of acting when issues arise are all needed. Utilizing literature from fields including AI Ethics, Sociotechnical Accountability, Auditing, Trust, Global Governance, and Education, this conceptual piece identifies the TRUTHIFY framework. The TRUTHIFY framework identifies six governance capabilities that need to be present to enable warranted trust in AI. They include: verifiable evidence; contestability and redress; independent assurance; aligned incentives and accountable ownership; inclusive capacity and participation; and adaptive coordination and AI literacy. All six capabilities are enacted at multiple levels (i.e., user-level and organization-level). Moreover, their ability to function effectively will depend upon a number of exogenous factors such as market pressures, degree of market concentration, geopolitical competition, institutional capacity, and specific risk characteristics of each sector. Finally, five propositions detail how TRUTHIFY's capabilities may facilitate the development of action-oriented practices that embody warranted trust. The paper further illustrates the application of TRUTHIFY's capabilities using Generative AI in Higher Education. This setting shows how TRUTHIFY can inform decisions about procurement, deployment, evaluation and review. The article offers an integrative model for assessing whether claims of responsible AI can be examined, challenged and enforced. Its broader aim is to clarify the conditions under which AI governance can support legitimate and credible reliance on AI systems.

Keywords: AI governance; warranted trust; contestability; AI auditing; generative AI; higher education

1. Introduction

AI systems can provide a wide range of services that benefit individuals and society. At the same time, they can create serious risks. One of the main challenges for researchers and policymakers is therefore deciding when these systems can be trusted and under what conditions that trust is justified. This paper addresses that problem by developing a conceptual governance framework called TRUTHIFY. The framework is illustrated through the case of generative AI in higher education. The paper also considers its wider implications for the development of legitimate AI innovation.

The goal of TRUTHIFY is to evaluate if organizations can demonstrate that their claims regarding responsible behavior are based upon actual practice.

TRUTHIFY evaluates whether an organization has demonstrated six institutional capabilities:

1. Verifiable evidence;
2. Contestability and redress;
3. Independent assurance;
4. Aligned incentives and accountable ownership;
5. Inclusive capacity and participation; and
6. Adaptive coordination and AI literacy.

These six capabilities are described in greater detail within Section Five of this paper. Collectively, they provide a mechanism to evaluate whether claims of responsible behavior regarding the deployment of artificial intelligence systems are supported by evidence, or merely rely on stated intentions.

This paper offers three primary contributions:

Firstly, it provides a connection between research regarding trustworthy AI, accountability, and global AI governance, all of which tend to be treated as distinct fields of study despite being linked through similar issues and challenges.

Secondly, it treats institutional incentives and institutional capability as critical factors in determining the success of governance models. Rather than viewing incentives and capacity as secondary elements, it views them as determinants of whether rules are followed, whether oversight processes are successful, and whether accountability is enforceable.

Thirdly, it highlights the importance of creating institutions that provide verifiable evidence of compliance through transparent processes that provide oversight and measurable forms of accountability.

Although the development of international AI governance has progressed significantly since the beginning of the last decade, gaps remain. While UNESCO's Recommendation on the Ethics of Artificial Intelligence has provided internationally agreed upon principles for ethics of AI such as respect for human rights, fairness, transparency, accountability, human oversight, inclusion and international cooperation (UNESCO, 2021), and the United Nations High Level Advisory Body on Artificial Intelligence has encouraged a more inclusive and interconnected approach to AI governance (United Nations, 2024); and although the UN General Assembly has established a Global Dialogue on AI Governance to promote dialogue amongst governments and other stakeholders (United Nations General Assembly, 2025), there is still significant room for improvement.

The broader aim is to support legitimate AI innovation. In this paper, legitimacy does not mean that an innovation is free from risk. It means that its development and use can be documented, questioned, reviewed, corrected and regulated by institutions that are answerable to those affected by its consequences. The paper addresses the following research questions:

RQ1. Why do convergent AI ethics principles often fail to produce warranted trust in real-world AI deployments?

RQ2. Which institutional capabilities translate high-level ethical commitments into verifiable, contestable and enforceable governance?

RQ3. How can these capabilities be coordinated across organizations and jurisdictions without reproducing digital dependency, exclusion or misplaced trust?

2. Conceptual Approach and Scope

This is a theory-synthesis conceptual article rather than an empirical study or systematic review. Conceptual papers can contribute by integrating previously separate literatures, clarifying constructs, explaining relationships and generating propositions that can later be tested (Jaakkola, 2020; Whetten, 1989). The problem-driven method used here began with a tension: ethical principles and governance instruments have multiplied, but justified trust remains uncertain. The analysis then connected five bodies of work that illuminate different parts of that tension: comparative AI ethics; sociotechnical accountability and transparency; organizational auditing and impact assessment; human trust in AI; and global inclusion, capacity and education.

Sources were selected purposively for conceptual relevance. The foundation includes comparative studies of AI ethics guidelines, established critiques of abstraction and transparency, peer-reviewed frameworks for auditing and reviewability and scholarship distinguishing trust from trustworthiness. It also includes authoritative governance instruments from UNESCO, the United Nations, the Council of Europe, the European Union and the U.S. National Institute of Standards and Technology. Recent education guidelines and content provenance are also included in this document as examples of how general principles are being applied within different sectors. This approach follows what Snyder (2019) recommends for an integrated literature review which is to clearly provide evidence of the purpose and rationale behind selecting the studies reviewed in order to avoid a misleading assertion of comprehensive review.
The unit of analysis in this study is the combination of the sociotechnical system and artificial intelligence; not just the individual model. Examples of elements which define the sociotechnical systems include data, users, models, interfaces, routines, purchasing agreements, policy decisions and communities impacted. In addition, the proposed framework will not consider technical performance separate from social context. An extremely accurate model could potentially be non-trustworthy when humans have no means to dispute its usage, relevant documentation does not exist, or there are no actors with the required authority and resource capacity to address potential harm. On the other hand, institutions may properly reduce their reliance upon an inaccurate model based on defining the intended purposes of such a model's application, tracking errors and maintaining meaningful human decision-making.

The article uses the term generative AI broadly for systems that produce text, images, audio, code, or other content in response to prompts. It does not attempt to resolve every frontier-safety issue, nor does it claim that a single framework can harmonize all legal systems. Its narrower question is institutional: what capabilities make governance commitments credible enough to justify context-specific reliance? The answer is developed by identifying recurrent limitations in the literature and then synthesizing them into a multi-level framework and propositions.

3. Literature Review

3.1 Convergence of ethical principles, divergence of implementation

There is a growing consensus in comparative analysis of AI ethics literature regarding the commonalities found in the many different types of ethical commitments made in AI ethics. For example, Jobin et al. (2019), identified five main recurring commitments to transparency, justice/fairness, respect for persons/non-maleficence, responsibility/accountability and

privacy/autonomy, as well as other less commonly mentioned ones, that were consistent with an array of guidance issued by governments, companies, professional organizations and civil society groups. Similar findings have been reported by Floridi et al. (2018), who identified a number of shared ethical principles for the development of trustworthy autonomous systems; namely, beneficence, non-maleficence, autonomy, justice and explicability. Cath et al. (2018), studied the conceptualization of AI's relationship to the public good by jurisdictions around the world. UNESCO has taken many of these common themes and transformed them into policy issues such as assessing impacts of AI, developing governance models for data generated by AI, gender and AI, environmental impacts of AI, using AI to improve education, improving healthcare through use of AI and the role of international cooperation in regulating AI. The fact that there exists a common language provides a foundation for reducing disagreements over whether human rights, accountability and inclusion should apply to AI.

The issue is that while agreement may exist at the abstract or high-level of discussion it may still hide important differences in terms of scope, trade offs, accountability and enforcement. High-level principles are insufficient to create the institutional settings, methodological standards and accountability mechanisms required for successful implementations. Many tools provide guidance on how to translate general principles into specific actions, however, there continues to exist fragmentation in this area. A fairness principle does not specify which type of fairness is applicable, whose harm counts, what constitutes adequate evidence or which actor(s) will take action when competing objectives arise. While principles can direct attention toward certain problems; they do not replace decision-making processes related to determining authority, procedures, available resources and potential consequences

3.2 Sociotechnical limits of transparency

Transparency is commonly presented as a route to accountability, yet disclosure is neither automatically intelligible nor empowering. Ananny and Crawford (2018) caution that seeing a system does not necessarily mean knowing how power operates through it. More information can leave core institutional relationships untouched, especially where the recipient lacks expertise, time, standing, or leverage. Selbst et al. (2019) identify “abstraction traps” that arise when a technical framing removes a system from the social context in which fairness and harm are produced. A model explanation may describe feature importance while saying little about why the system was introduced, whether the task should be automated, how the decision is reviewed, or which institutional alternatives were rejected.

Documentation practices address the institutional translation problem by making development assumptions and limits more visible. For example, datasheets for datasets and model cards provide standardized reporting requirements for provenance, intended use, performance and limitations (Mitchell et al., 2019; Gebru et al., 2021). Content-provenance and watermarking standards attempt to document where content originates and how it is altered. As a result, users are able to assess better whether content is trustworthy. However, these mechanisms increase the quality of evidence available. Their governance value still depends on who is required to generate the documentation, whether claims made based on such documentation may be verified independently and what happens when such documentation is incomplete or misleading.

3.3 Accountability, reviewability and assurance

Accountability requires more than visibility. Wieringa (2020) describes algorithmic accountability through relationships among actors, forums, standards, information, debate, judgment and

consequences. This relational approach is important because an organization is not accountable in the abstract; it is accountable to a forum with the authority and capacity to question conduct and impose consequences. Kroll et al. (2017) similarly argue that accountable algorithms require procedural and institutional mechanisms, not only access to source code. The relevant question is whether a decision process can be reconstructed and evaluated against applicable legal and organizational standards. Cobbe et al. (2021) develop reviewability as a way to structure records across the full decision process, including the organizational choices that precede a model output and the actions that follow it. Reviewability requires the creation of an ongoing record of evidence. Assessments of impact identify what could go wrong either before or at the time of implementation. However, Metcalf et al. (2021) indicate that impacts are identified as much as they are created through choices related to stakeholders, categorization, methodology and institutional boundaries. Therefore, an assessment may be designed to exclude harmful outcomes while providing the appearance of comprehensiveness.
Frameworks for audits aim to determine if systems have met established standards via some form of independent or organized review. Raji et al. (2020) present an example of an internal audit process for large language models that ties together stages of organization development with documentation and accountability. Mökander et al. (2021) note that ethics based auditing offers a promising form of governance with significant limitations, including concerns regarding the independence of auditors, their role, standards and access. A multi-layered approach can be used for evaluating large language models by examining the model itself, the environment within which it is deployed and the use case (Mökander et al., 2024), making governance a more tangible entity. However, until an audit's findings are presented to those who make decisions on the basis of the findings, lead to actions to address issues found and can influence approval, procurement, liability, etc. -- or continued deployment -- an audit will remain uncredible.

3.4 Trust, trustworthiness and calibration

Trust in systems has two forms -- institutional and psychological. Studies on trust in Artificial Intelligence (AI), reviewed by Glikson & Woolley (2020), show that performance, transparency, tangibility, task type and Human-Like features influence trust. However, these factors do not necessarily reflect actual system performance. For example, a user-friendly conversational interface could increase the user's confidence in the system even if the output of the interface was inaccurate. Therefore, Jacovi et al. (2021) make the distinction between warranted trust, which exists based on trustworthiness; and unwarranted trust, which does not exist based on trustworthiness. Warranted trust changes what we view as a "good" outcome from achieving positive sentiments toward an AI System, to having a balance between reliance on the system and evidence that warrants the reliance.

Toreini et al. (2020) relate three social science-based aspects of trust -- ability, benevolence and integrity -- to three technical aspects of trust -- Fairness, Explainability/Auditability/Safety. They emphasize that trustworthiness relies upon multiple stakeholders/actors at multiple stages of a system. This provides strong justification for developing an enterprise-wide or systems perspective on creating trustworthy AI systems. However, this also creates institutional concerns about how to provide assurance that a system is fair, explainable and safe, including providing a mechanism for evaluating contestability and accountability.

3.5 Global governance, dependency and capacity

The "cross-border" aspect of AI supply chains and their influence is an area in which "legal authority", "infrastructure", and "regulatory capability" are still unevenly distributed across national borders. In response to this, Gasser and Almeida (2017) have proposed a multi-layered governance structure that addresses three levels of governance; namely the technical, ethical, and legal. In addition to suggesting a multi-level governance structure, Cihon et al. (2020) explored possible alternatives to existing international governance structures regarding AI and noted that if such alternative models become fragmented, they could lead to the development of inconsistent or competing models of regulation. The UN has issued a call for a "globally inclusive and distributed governance architecture" for AI, as opposed to one dominated by a few powerful states and corporations (United Nations, 2024). A global dialogue on AI governance is currently being convened through the General Assembly's ongoing Global Dialogue on AI Governance (United Nations General Assembly, 2025). This will provide an opportunity for participants to discuss issues related to the need for inter-operable standards, capacity building, human rights protection, transparency and oversight. However, formal recognition of all stakeholders does not resolve underlying structural inequalities. Using decolonial theory, Mohamed et al. (2020), examined how AI technologies can reinforce extractive relationships

with respect to data, labor, knowledge and institutional power. Birhane (2020), described the risks associated with what she termed 'algorithmic colonization' in Africa. These include imposition of foreign values and dependency upon foreign-developed systems. Sambasivan et al., (2021) explain why fairness concepts cannot simply be translated into local contexts based upon abstract measures, but must be re-imagined in light of specific local social norms and histories. These critiques imply that global governance must distribute agenda-setting power, technical resources and evaluative authority, not merely invite participation after standards have been defined.

3.6 Education, literacy and the provenance of knowledge

Generative A.I. can assist with explanations, translations, accessibility, feedback and developing ideas – but it can also provide misrepresented information, cause a loss of privacy, help produce false evidence and replace the outcome of an individual's learning process with the actual output of the A.I. itself. Cotton et al. (2024), authors who have studied responses to Academic Integrity related to A.I. use in Higher Education suggest that such responses will require assessing the way we evaluate assignments and establishing clear expectations as opposed to relying solely upon detecting cheating. UNESCO suggests that a human-centered approach that preserves the agency, privacy, inclusiveness and age-appropriate use of individuals is preferable to an approach that relies upon compliance-based regulations (UNESCO, 2023, 2024a, b). Additionally, UNESCO provides competency frameworks that address A.I. literacy through a combination of human-centered values, ethics, techniques, critical judgement and responsible design (UNESCO, 2023, 2024b).

4. Research Gap

The literature identifies most of the ingredients needed for responsible AI governance, but they remain divided across distinct problem frames. Ethics-guideline research explains convergence and vagueness. Transparency scholarship explains why disclosure can fail. Accountability research offers forums, reviewability, impact assessment and auditing. Global governance and decolonial scholarship expose unfair access to capacity and agenda-setting power. Educational guidelines associate institutional safeguards with literacy and human agency. Less developed is an account of how these elements work collectively as institutional capabilities that convert principles into credible commitments.

Three specific gaps follow. First, many frameworks move from principles to tools but do not make production of warranted trust its express dependent outcome. Therefore, organizations may count disclosures, assessments, training sessions without determining whether reliance is actually supported by reviewable evidence and realistic remedies. Secondly, governing frameworks usually treat competition, concentration, geopolitical rivalry and organizational incentives as context which weakens implementation. Voluntary audit or code will work under cooperative conditions but fail when market entry, procurement deadlines, strategic advantage reward speed and secrecy. Thirdly, global inclusion is frequently stated as principle but rarely integrated into the operating design of assurance, redress and standards. Participation without resources, language access or decision authority can legitimize arrangements that preserve dependency. The conceptual opportunity is therefore not to propose another universal list of ethics principles. It is to specify a linked set of institutional capabilities and explain the mechanisms through which they make responsible-AI claims verifiable, challengeable, consequential, inclusive and adaptable. The framework must operate across levels because a user's ability to question an output depends on organizational records, while organizational behavior depends on law, standards, market incentives and cross-border infrastructure. It must also recognize that trust can be excessive or deficient. The desired outcome is calibrated, warranted trust, accompanied by justified non-reliance where evidence or safeguards are inadequate.

Table 1. Literature problems and the institutional function of TRUTHIFY capabilities

Literature problem	TRUTHIFY capability	Core governance question	Selected foundations
Principles are broad and implementation is fragmented	All six capabilities	What institutional arrangements convert a principle into a repeatable practice?	Jobin et al. (2019); Mittelstadt (2019); Morley et al. (2020)
Transparency can disclose without empowering	Verifiable evidence; contestability and redress	Can relevant actors check a claim and act on the information?	Ananny and Crawford (2018); Cobbe et al. (2021)

Audits and assessments may be narrow or symbolic	Independent assurance; aligned incentives	Are reviewers independent and do findings change decisions or produce consequences?	Raji et al. (2020); Metcalf et al. (2021); Mökander et al. (2021)
Human confidence may exceed actual trustworthiness	Evidence; assurance; literacy	Is reliance calibrated to performance, limits and institutional safeguards?	Glikson and Woolley (2020); Jacovi et al. (2021); Toreini et al. (2020)
Global governance can reproduce dependency	Inclusive capacity and participation; adaptive coordination	Do affected communities and lower-resource institutions have influence, expertise and infrastructure?	Birhane (2020); Mohamed et al. (2020); Cihon et al. (2020)

Source: Authors' own elaboration.

5. The TRUTHIFY Institutional Trust Framework

TRUTHIFY treats AI governance as an institutional system for producing and testing reasons for reliance. Its inputs are ethical principles, legal duties, professional norms, technical standards and organizational objectives. These inputs become credible through six capabilities. The capabilities create five mechanisms - information quality, procedural justice, credible commitment, distributed legitimacy and organizational learning - that support the outcome of warranted trust and legitimate innovation. The relationship is moderated by sectoral risk, market concentration, competitive pressure, geopolitical rivalry and institutional capacity. Figure 1 presents the model.

5.1 Capability 1: Verifiable evidence

Verifiable evidence concerns the records and technical signals needed to evaluate an AI system and its outputs. It includes documentation of data sources, model limitations, intended uses, evaluation results, change histories, human decisions and content provenance. Evidence must be proportionate to risk and available to the actors who need it, including regulators, auditors, organizational reviewers, affected people and researchers under appropriate confidentiality controls. Model cards and datasheets illustrate structured documentation, while provenance standards demonstrate how claims about digital assets can be bound to tamper-evident records (Coalition for Content Provenance and Authenticity, 2026; Gebru et al., 2021; Mitchell et al., 2019).

Evidence does not necessarily mean open evidence. There are many reasons why one may have tiered access (e.g., security issues, privacy concerns, trade secrets or the risk of abuse). The important point is that there needs to be an opportunity for a sufficient third-party expert who has competence to independently check the consequential claims made by the institution. A trustworthy institution would be able to respond as follows about their use of an automated system: What automated system did you use? Why did you use it? In what versions and with what data practices? Were there any recognized limitations in how the system performed? Was the

performance of the automated system adequately evaluated within your target population and/or language(s)? Who made the human decisions that influenced the outcome of the use of the automated system?

5.2 Capability 2: Contestability and redress

Transparency without Contestability is Inadequate. Contestability enables impacted individuals to challenge the use of AI in supporting their case(s), provide evidence related to their claims, and get reviewed and/or corrected/compensated. Contestability requires formal notification of AI use, a logical basis for how AI was applied (with respect to its application) as well as multiple avenues for access, specific timeframes for responses, protection against retaliation and a pathway to appeal to either a human being or an independent body.

Redress can include corrections to the record; reconsideration of previous decisions; monetary damages if appropriate; as well as adjustments to the overall AI systems with patterns of harm are found. This will convert transparency into actual power through procedure. The ability to act on disclosed information is limited by the ability to do so. Contestability provides support for reviewability by maintaining records throughout technical and organizationally based decision-making processes (Cobbe et al., 2021). If a process cannot produce enough evidence for meaningful review, the organization should narrow the use, retain stronger human control, or avoid automation. In this sense, appeal is not an afterthought; it is a design requirement that disciplines what can responsibly be automated.

5.3 Capability 3: Independent assurance

Independent assurance assesses claims regarding the safety, equity/fairness, security/rights impacts/documentation/governance/ etc. of systems. Independent assurance can take numerous forms such as internal audit units with protected independence, external auditors, regulatory testing/civil-society evaluations/red teams/conformity assessments/public reporting. The type of independent assurance required will depend upon the level of risk involved and the applicable

laws/regulations. In cases where high-risk use is anticipated greater separation between those who develop/purchase systems and those who independently assure their acceptability will typically be needed.

Assurance should examine the model, the surrounding system and the application context. Model-level testing can miss harms caused by workflow, interface, policy, or human incentives. System-level review can miss unequal effects in a particular school, language community, or public service. Layered auditing responds to these limitations, but assurance remains vulnerable to narrow scope, controlled data access, auditor dependence and non-public findings (Mökander et al., 2021, 2024; Raji et al., 2020). The capability is therefore defined not by the existence of an audit, but by credible independence, access to evidence, relevant expertise and a documented route from findings to action.

5.4 Capability 4: Aligned incentives and accountable ownership

The capability to assign accountable owners, decision rights, resources and consequences is necessary to make governance last beyond the current administrative term. Included in this capability is board or senior management responsibility; procurement conditions; contractual audit rights; protected budgets for oversight; and consequences for evasion. If there are no incentives within an organization to comply with governance requirements, then documentation, assurance and participation will likely be viewed as costs to minimize rather than obligations to meet.

5.5 Capability 5: Inclusive capacity and participation

In order to effectively establish inclusive governance, there needs to be more than just a process of consulting with stakeholders. There must be a capacity to know how to utilize an understanding of AI systems so as to both shape and challenge them. The inclusion of affected communities in agenda setting; the provision of linguistic and accessible support for participation; the creation of funding streams for public-interest experts; the development of regulatory and technological capabilities in less affluent areas; the establishment of mechanisms to recognize local knowledge; and, most importantly, the inclusion of participation at the time of establishing purposes and standards as opposed to providing comment after completion of a model are all essential components of this capability.

5.6 Capability 6: Adaptive coordination and AI literacy

Adaptive governance of AI systems, their uses, and risks will have to take place over time. As such, governance will likely need to involve continuous monitoring, incident learning, periodic assessments/reassessments, version control, and inter-organizational/jurisdictional coordination.

The mechanisms needed for adaptive coordination include (but are not limited to) interoperability in reporting, shared terminology, standards mapping, cross-jurisdictional incident reporting channels, research access, and regular policy reviews. While these mechanisms may allow for some degree of differences in law, they need sufficient levels of compatibility so that the evidence, assurances and redress mechanisms do not simply "disappear" as you move across organizational or national boundaries.

Similarly, AI literacy will also support this adaptive capability as both users, professionals/leaders, regulators and educators will each have role-based competences. AI literacy should enable people to identify uncertainty, authenticate source material, protect sensitive data/information, identify proper/prohibited use of the system(s), and provide the knowledge/means to dispute decisions. In no case can AI literacy be used to shift institutional responsibilities to individual actions. For example, if a student has an obligation to test/check answers then it is the University's responsibility to evaluate the tool being utilized by students. Similarly, while regulatory agencies can provide consumers with warnings about certain products/services, such warnings cannot serve as substitutes for remedies. Regulatory agency personnel training on new technologies does not offset their lack of access to technological information/evidence. When combined with institutional safeguards and periodically updated through experiences, AI literacy enables people to make informed decisions regarding when it is reasonable to rely upon a given technology.

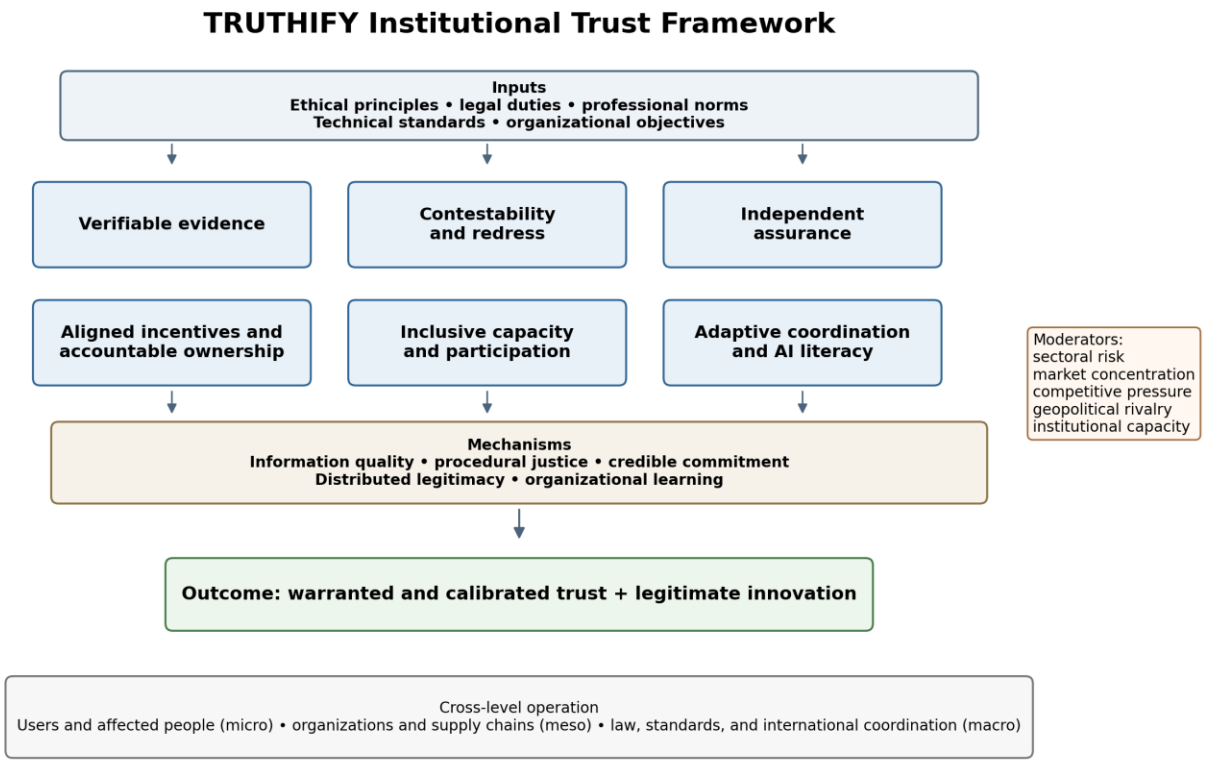
5.7 Cross-level dynamics and boundary conditions

The three interdependent levels of this system are as follows. Users and affected individuals interact with the outputs, explanations, messages, options for action and routes to appeal that make up the micro-level of the system. Organizations and supply-chains determine how they will acquire products, services or information; how they will document their activities; what testing is required; who gets to determine the flow of work; who gets to decide on incentives and whose responsibility

is to provide remediation at the organizational/supply-chain (meso) level. Standards organizations, professional associations, civil society, international institutions and governments create and enforce the minimum responsibilities of organizations in relation to each other (macro). Each level is dependent upon the success of the others. An informed consumer cannot get redress if there was no documentation by the organization. An organization may be unable to assess whether a large vendor is acting appropriately if the contract does not allow it access to the vendors' records. No regulatory body will be able to require it to do so either.

Existing governance instruments can be interpreted as partial supports for these levels. NIST's AI Risk Management Framework organizes governance, mapping, measurement and management across the system life cycle, while its generative AI profile identifies risks and actions specific to generative systems (National Institute of Standards and Technology, 2023, 2024). The European Union's AI Act creates risk-based legal duties and the Council of Europe's Framework Convention links AI governance to human rights, democracy and the rule of law (Council of Europe, 2024; European Parliament and Council of the European Union, 2024). TRUTHIFY does not replace these instruments. It provides a common analytical lens for asking which capabilities they create, for whom and with what evidence and consequences.

TRUTHIFY's multi-level moderation logic adds to these instruments. TRUTHIFY models how capability at one level impacts capability at other levels. It also models how sector risk, market concentration, competitive pressure, geopolitical rivalry and institutional capacity affect the threshold level of capability required. Therefore, it is diagnostic as opposed to prescriptive. Boundary conditions affect how strong each capability must be. Higher sectoral risk increases the need for independent assurance, formal contestability and enforceable ownership. Market concentration can weaken procurement leverage and make shared regulatory or public-interest testing more important. Competitive pressures may reduce the time required to complete a review of new uses and provide an incentive for companies to disclose only the information needed by regulators. Geopolitics may limit international collaboration or allow countries to claim national security as a basis for restricting cooperative efforts. If regulatory bodies do not have sufficient resources (i.e., institutional capacity) to enforce their rules, even well-written rules will become mere "paper" regulations. This framework predicts that the weakest link in the chain of governance will occur when high-risk uses are combined with concentrated supply; strong economic incentives to deploy new technologies quickly; and weak institutional capacity to oversee new uses. Until such time as robust safeguards can be developed, all parties should rely on each other less, set more stringent evidentiary requirements before taking action, and develop more detailed plans for how they intend to respond to risks associated with new uses.



Source: Authors' own elaboration.

Figure 1. TRUTHIFY institutional trust framework. The model treats warranted trust as an outcome of interacting capabilities rather than a direct result of ethical principles.

6. Conceptual Mechanisms and Propositions

The six capabilities are interdependent. Evidence verifiably proves the information, yet this evidence is made behaviourally relevant (i.e. usable) through contestability and assurance. Contestability provides procedural justice; however, contestability relies on both records and a forum that has authority. Assurance tests claims; meanwhile, alignment of incentives provides consequence to findings. Adaptive coordination transforms incidents and new evidence into modified practice. Likewise, inclusive capability enlarges the scope of interests and knowledge used within evaluation. Together, the potential mechanisms create justified trust: reliance proportional to the amount of warranted evidence and protective measures. Proposition 1: There will be a greater positive effect on warranted trust related to transparency and documentation as warranted evidence disclosure is correlated with substantive contestability and

review processes versus evidence disclosure alone.

This proposition results directly from the limitations of transparency. While information can provide improved decision-making, only those who can understand the information, dispute its relevance, and receive a response regarding their concerns can use the information effectively. A model card placed on-line may aid researchers, but likely does nothing for someone impacted by an automated decision. Similarly, an appeals process that does not require records may generate some type of formal review; nonetheless, there would still be no reliable means to evaluate whether judgments were correct. When combined together, information is transformed into procedural capacity.

This expectation is supported through literature that documents that disclosure alone will not lead to accountability where impacted parties have no ability to raise questions regarding what was disclosed (Ananny & Crawford, 2018; Cobbe et al., 2021).

Proposition 2: Independent assurance will strengthen warranted trust only to the extent that audit findings are linked to remediation, deployment decisions and consequences for accountable owners.

The potential impact of assurance will depend upon the path through which it has been instituted within an organization. That is, in order for assurance to have its anticipated effects, findings must be conveyed to individuals or groups having authority over the system or process being assured. Further, such findings must result in a documented response. Finally, those findings must influence decisions regarding the launch, change, restriction or withdrawal of the system or process.

Literature on existing accountability also provides support for this conditional perspective, indicating that audit and assessment of impacts may become part of routine business processes and still fail to provide meaningful changes in outcomes due to scope and response remaining under the control of audited entities (Metcalf et al., 2021; Raji et al., 2020).

Proposition 3: Incentive alignment will moderate the relationship between formal AI governance commitments and responsible organizational behavior, particularly under high competitive or geopolitical pressure.

Where speed, market share, cost reduction, or strategic advantage dominate performance assessment, broad ethical obligations are likely to be diminished. Ownership clearly defined, clear procurement rules, liability, protected oversight and enforcement credible enough for organizations to protect themselves against evasion of their own actions and strengthen the

organizational status of governance work in no way implies that punitive regulation is always the best course of action. Rather, it indicates that responsible behavior must be supported by decision making processes and incentives powerful enough to withstand the conflicts with short-term goals.

This supports translation into practice as a conditional process in which pressures for performance override general principles unless there are governance processes that cause ethical commitments to have consequences (Mittelstadt, 2019; Morley et al., 2020).

Proposition 4: Inclusive capacity and participation will increase the legitimacy and contextuality of AI governance when participants have influence over agendas, necessary resources, and access to evidence.

Capacity can change the level of participation. When communities and jurisdictions with fewer resources can provide their own expertise, language, funding, and position/standing in relation to harms created by imported technology, they will also be able to identify the harms that were left out of consideration, challenge the same foreign assumptions and create alternative solutions. Therefore Proposition 4 asserts that participation based upon real influence will improve both perceived legitimacy and fit to purpose.

Studies concerning how low-income populations are treated fairly regarding algorithms show that engaging these populations actively in the development and deployment of algorithms while failing to provide the appropriate resources and position/standing and access to evidence tends to reinforce those original foreign assumptions instead of correct them (Mohamed et al., 2020; Sambasivan et al., 2021).

Proposition 5: Embedding AI literacy within verifiable, contestable and accountable institutional mechanisms will enhance trust calibration; using literacy alone may produce small or negative effects or place too much accountability on the user.

Education can assist individuals in recognizing uncertainty and use AI appropriately. Education cannot grant an individual access to a vendors training data, security testing or internal incident records. Therefore, literacy is supplementary to institutional capabilities. Well-governed environments allow informed users to read through disclosures, notice issues and initiate review processes. Poorly-governed environments can disguise the duties of developers, deployers and regulatory bodies behind recommendations for 'users' to 'use AI responsibly'.

This is consistent with trust research that distinguishes warranted from unwarranted reliance and indicates that calibration occurs based on the characteristics of a system and its governance structure and not solely based upon user intentions (Toreini et al., 2020; Jacovi et al., 2021).

These propositions are not claims of empirical validation. They specify relationships that can guide comparative case studies, surveys, field experiments, audit research and policy evaluation. They also identify failure modes. An arrangement may be transparent but not contestable, audited but not consequential, participatory but not resourced, or literate but institutionally unprotected. Such partial arrangements can create the appearance of governance while leaving trust unwarranted.

7. Illustrative Application: Generative AI in Higher Education

A university has developed a Generative-AI Tutor to integrate with their Learning-Management System (LMS) which will include a variety of features such as explaining concepts, creating practice questions, translating materials and providing feedback. These tools are expected to increase student's ability to access tutoring for those who do not have access to tutors or need help with language and/or accessibility needs. However, there could be risks associated with this technology including: providing false explanations, exposing personal data, reproducing cultural bias, promoting dependency and possibly blur the line of distinction between tutoring/assistance and authorship. Therefore, the University must decide not only if the tool is innovative but also at what time(s) students and teachers can depend upon the technology.

Table 2 translates the application into practical questions. The example demonstrates why prohibition and unrestricted adoption are both inadequate. A ban can sacrifice accessibility and learning benefits while driving use underground. Unrestricted use can replace intellectual effort, privatize educational data and create false confidence. The framework supports bounded experimentation: clearly defined uses, strong evidence and review, human responsibility and the ability to withdraw reliance when safeguards fail.

Table 2. Applying TRUTHIFY to a university generative AI deployment

Capability	Higher-education implementation question	Illustrative evidence or control
Verifiable evidence	What is the tool, what data does it use and where is it reliable?	Version records; data-flow map; course-relevant testing; limitations; provenance/disclosure
Contestability and redress	How can students or staff challenge an output or AI-supported action?	Notice; human review; complaint channel; response times; correction and non-retaliation

Independent assurance	Who evaluates accuracy, privacy, security, bias, accessibility and pedagogy?	Pre-deployment review; periodic audit; red teaming; public summary; suspension authority
Aligned incentives and accountable ownership	Who can approve, pause and accept responsibility for the deployment?	Named owners; procurement clauses; audit rights; incident notice; exit rights; deployment gates
Inclusive capacity and participation	Who shapes purpose, rules and evaluation and who may be excluded?	Student/faculty participation; disability and language expertise; equitable alternatives; capacity support
Adaptive coordination and AI literacy	How will policy, competencies and controls change with evidence and model updates?	Role-based training; termly review; incident learning; version reassessment; standards mapping

In addition, a brief comparison to a different type of system illustrates that the proposed framework is not specific to education. For instance consider a public entity utilizing an artificial intelligence system to generate eligibility correspondence for benefits. The AI-generated correspondence would need to include information such as the prompt used to create the decision-making process, the version number of the AI model and all manual edits made to the AI-generated correspondence prior to issuing the letter. These items would constitute verifiable evidence. Contestability requires that a claimant can file a complaint and receive a response from someone authorized to rescind a determination. Independent assurance means that review of the documentation created to make eligibility determinations would require to be reviewed by a group outside of the office responsible for making those determinations. Accountable ownership refers to identifying a specific employee of the public entity responsible for accuracy of results. Inclusive capacity requires that claimants with limited literacy or limited English proficiency can use the appeals process. Finally adaptive coordination requires that if there are repetitive errors in generating eligibility correspondence then these issues should result in a modification to either the deployment of the AI system or at least the training materials provided to users.

8. Governance, Ethical and Practical Considerations

In order to select appropriate governance measures, organizations must first define the decision in question, groups that will be impacted, level of harm (and whether harm is reversible) and an acceptable level of residual risk.

An ethical concern that cuts across all other concerns is power asymmetry. For example, vendors may have control over the evidence of technology, while institutions may have control over the process to file an appeal. States may invoke national security and end-users may simply have limited options regarding their selection of technologies. The right to confidentiality can be valid; however, this does not mean that independent scrutiny should be eliminated from the process. Claims regarding trade secrets need to be weighed against the rights of regulators, auditors and courts to access information. All parties (those using AI, as well as those impacted by its non-use) should participate in the development of AI. This is true because refusing to implement AI can still impact the ability of individuals or communities to access resources, services, or opportunities.

However, the framework for developing and implementing AI will create costs. These costs include the documentation required to support evaluations, accessible mechanisms for appealing decisions related to AI implementation and capacity building. As such, proportionality is critical for smaller organizations and jurisdictions with fewer resources. Establishing shared testing

facilities, providing public interest technical assistance, creating common formats for documentation, pooling procurement efforts and developing interoperable standards can help reduce duplicative efforts. International cooperation is equitable only when international cooperation increases each jurisdiction's capabilities rather than establishing requirements based on an actor with existing infrastructure to meet those requirements.

Lastly, the framework allows for a role for democratic judgment. While evidence and audits can inform decisions about how to use AI, they cannot dictate what value systems will guide every decision. Decisions related to acceptable levels of surveillance for education purposes, replacement of jobs through automation and deployment of AI for military applications are questions of politics and ethics. TRUTHIFY can provide useful insight when these decisions become visible and subject to review and challenge; however, TRUTHIFY should not be employed to transform disputed value judgments into quantifiable scores.

9. Limitations

The four major disadvantages to this study include that (1) this research is conceptual and does not provide empirical testing of the proposed concepts. (2) The studies selected were not systematically searched therefore potentially relevant studies could have been excluded particularly from non-English language sources and/or those not indexed in major databases. (3) The six capabilities identified in the previous section overlap with one another. Documentation supports both Assurance and Contestability while Capacity can influence all of the other capability types. The categories provided here represent an analytical distinction as opposed to a separate departmental structure within an organization.

(4) The proposed framework identifies the institutional elements required to establish warranted trust however, it provides no comprehensive treatment of Technical Safety, Cybersecurity Engineering, Frontier Model Risk, Environmental Impact, Labor Conditions, or Military Uses. Additionally, the political economics of concentrated computing and platform power cannot be resolved using this framework. However these issues will be included in the model through evidence requirements, risk moderating variables and subject of accountability; each requiring additional investigation.

10. Future Research

The first step in future research will be to define measures for all six capabilities (verifiable evidence; contestability and redress; independent assurance; aligned incentives and accountable ownership; inclusive capacity and participation; and adaptive coordination and AI literacy). These may include:

Documentation: Completeness and Accessibility;

Appeals: Time to Appeal and Success Rates;

Auditing: Independence of Auditor and Access to Audit Results;

Remediation: Timely Closure of Remediated Issues;

Decision Authority: Allocation of Decision-Making Authority among Stakeholders;

Stakeholder Influence: Degree of Influence of Stakeholders on Decisions;

Local Language Evaluation: Whether Evaluations are Conducted in a Local Language;

Incident Sharing Quality: Effectiveness of Incident Sharing Among Developers and Deployers;

Changes in User Calibration: Changes in Users' Confidence in their Ability to Make Trustworthy Decisions.

The second step is to differentiate between Perceived Trust and Warranted Trust. This can be accomplished through comparisons of users' levels of confidence in their ability to make trustworthy decisions, as well as actual system performance and conditions of governance.

Comparative Case Studies

Case studies examining similarities and differences among various types of organizations (e.g., Universities, Public Agencies, Financial Institutions, Nonprofit Organizations, Technology Firms) would allow researchers to determine if the Framework provides explanations for observed differences in organizational practices related to assurance.

Research in Lower-Resource Settings

In addition to conducting research in developed countries, researchers should also conduct research in developing countries. It is recommended that this type of research be conducted jointly with local institutions so that researchers can better understand the effects that International Standards for assurance have on building institutional capacity or creating dependency.

Central Empirical Question

The key empirical question here is not "Does an Organization Use Ethical Language?" but rather "Do the Arrangements made by an Organization Provide Reliable Reasons for Trust, Reliable Reasons to Withhold Trust and Effective Routes to Correctness?

11. Conclusion

Ultimately the main governance problem for generative artificial intelligence (A.I.) is not only how to encourage innovation in developing these systems, nor how to contain the potential risks of generating A.I., but how to develop institutions that enable responsibility claims regarding A.I. to be considered trustworthy when competing under uncertain conditions.

Ethical principles are necessary; however, ethical principles cannot implement themselves. Transparency can provide information – yet empower no one. Audits can take place – yet audits will have no consequence. People can participate – yet people participating will have no voice/influence. Literacy can be required/mandated – yet the lack of literacy does not protect anyone.

TRUTHIFY addresses this by treating warranted trust as an institutional achievement. In order to achieve trusted relationships among parties using verifiable evidence, creating mechanisms for contestability and redress, having third party independent assurances that actions taken are in alignment with incentives and ownership structures, establishing mechanisms for inclusivity and participation and coordinating efforts adaptively while providing literacy support; all of these factors work together in a complex system to ensure information quality, procedural justice, credible commitments, distributed legitimacy and ongoing learning. TRUTHIFY's contribution is intentionally modest and simply provides a model that connects existing knowledge on how to

design effective governance frameworks that ask if your governance framework is able to be reviewed, has consequences, is inclusive and adaptable.

For Supervisors, Reviewers, Policymakers and Practitioners there are very simple ways to assess whether reliance on AI should be made. There are four questions you need to answer: Who can confirm why I used this A.I.? Who can dispute my use of this A.I.? Who will be held accountable if this A.I. fails? And did I have the ability to determine the structure/rule for the development/use of this A.I.?

References

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973-989. https://doi.org/10.1177/1461444816676645

Birhane, A. (2020). Algorithmic colonization of Africa. *SCRIPTed*, 17(2), 389-409. https://doi.org/10.2966/scrip.170220.389

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial intelligence and the 'good society': The US, EU and UK approach. *Science and Engineering Ethics*, 24(2), 505-528. https://doi.org/10.1007/s11948-017-9901-7

Cihon, P., Maas, M. M., & Kemp, L. (2020). Fragmentation and the future: Investigating architectures for international AI governance. *Global Policy*, 11(5), 545-556. https://doi.org/10.1111/1758-5899.12890

Coalition for Content Provenance and Authenticity. (2026). *C2PA technical specification (Version 2.4)*. https://spec.c2pa.org/specifications/specifications/2.4/index.html

Cobbe, J., Lee, M. S. A., & Singh, J. (2021). Reviewable automated decision-making: A framework for accountable algorithmic systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 598-609). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445921

Council of Europe. (2024). *Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law* (CETS No. 225). https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=225

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 61(2), 228-239. https://doi.org/10.1080/14703297.2023.2190148

European Parliament and Council of the European Union. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence*. Official Journal of the European Union, L., 2024/1689. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). *AI4People - An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations*. *Minds and Machines*, 28(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5

Gasser, U., & Almeida, V. A. F. (2017). A layered model for AI governance. *IEEE Internet Computing*, 21(6), 58-62. https://doi.org/10.1109/MIC.2017.4180835

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daume III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86-92. https://doi.org/10.1145/3458723

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2), 627-660. https://doi.org/10.5465/annals.2018.0057

Jaakkola, E. (2020). Designing conceptual articles: Four approaches. *AMS Review*, 10, 18-26. https://doi.org/10.1007/s13162-020-00161-0

Jacovi, A., Marasovic, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 624-635). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445923

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1, 389-399. https://doi.org/10.1038/s42256-019-0088-2

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, 165(3), 633-705. https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3/

Metcalf, J., Moss, E., Watkins, E. A., Singh, R., & Elish, M. C. (2021). Algorithmic impact assessments and accountability: The co-construction of impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 735-746). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445935

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1, 501-507. https://doi.org/10.1038/s42256-019-0114-4

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220-229). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596

Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33, 659-684. https://doi.org/10.1007/s13347-020-00405-8

Mökander, J., Axente, M., Casolari, F., & Floridi, L. (2021). Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations. *Science and Engineering Ethics*, 27, Article 44. https://doi.org/10.1007/s11948-021-00319-4

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2024). Auditing large language models: A three-layered approach. *AI and Ethics*, 4, 1085-1115. https://doi.org/10.1007/s43681-023-00289-2

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, 26, 2141-2168. https://doi.org/10.1007/s11948-019-00165-5

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

National Institute of Standards and Technology. (2024). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (NIST AI 600-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.600-1

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. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33-44). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372873

Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T., & Prabhakaran, V. (2021). Re-imagining algorithmic fairness in India and beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 315-328). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445896

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59-68). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287598

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104, 333-339. https://doi.org/10.1016/j.jbusres.2019.07.039

Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Gonzalez Zelaya, C., & van Moorsel, A. (2020). The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 272-283). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372834

United Nations. (2024). *Governing AI for humanity: Final report of the High-level Advisory Body on Artificial Intelligence*. https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf

United Nations General Assembly. (2025). *Terms of reference and modalities for the establishment and functioning of the Independent International Scientific Panel on Artificial Intelligence and the Global Dialogue on Artificial Intelligence Governance* (A/RES/79/325). United Nations. https://digitallibrary.un.org/record/4087699

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. https://unesdoc.unesco.org/ark:/48223/pf0000381137

UNESCO. (2023). *Guidance for generative AI in education and research*. https://unesdoc.unesco.org/ark:/48223/pf0000386693

UNESCO. (2024a). *AI competency framework for students*. https://unesdoc.unesco.org/ark:/48223/pf0000391105

UNESCO. (2024b). *AI competency framework for teachers*. https://unesdoc.unesco.org/ark:/48223/pf0000391104

Whetten, D. A. (1989). What constitutes a theoretical contribution? *Academy of Management Review*, 14(4), 490-495. https://doi.org/10.5465/amr.1989.4308371

Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 1-18). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372833