

From transparency to accountability in AI: a conceptual synthesis

Conceptual paper · Prepared October 21, 2026

Prepared as a product example with AI assistance. This selective synthesis is not a systematic review, a published study, an independent human review, or an approved research run.

Abstract

Transparency can make an AI system available for inspection without specifying who must respond to what the inspection reveals. This conceptual synthesis examines that distinction through five selected foundational works on AI ethics, transparency, and auditing. It proposes a relationship between three activities: making relevant information accessible, enabling an audience to interpret and challenge that information, and assigning responsibility for a response. The argument is that disclosure supports accountability when these activities connect around a defined decision. Three propositions specify the roles of audience, response authority, and continuing documentation. They are offered as analytical suggestions rather than tested causal findings. A hypothetical organizational example illustrates how the framework could guide questions about a deployment decision. The paper does not establish the prevalence or effectiveness of any governance practice. Its contribution is a bounded vocabulary for examining whether a transparency arrangement provides a usable route from information to answerability and correction.

Introduction

A practical question motivates this paper: what must happen after information about an AI system becomes visible? The question distinguishes the production of an explanation from the allocation of responsibility for responding to it. Consider an organization that publishes an account of a model's intended use. That account might help a reader understand a decision, yet it does not by itself identify a person who must answer a challenge. The present argument treats this possible gap as a problem of relationships among information, audiences, and institutional action. The distinction is analytical. It is not evidence that disclosure always fails or that a particular accountability mechanism necessarily succeeds.

Jobin, Ienca, and Vayena identify recurring ethical principles across the guidelines they examine while reporting substantial differences in interpretation and implementation. Their analysis supplies a reason to distinguish agreement on a term from agreement on its practical requirements. Hagendorff examines ethical guidelines alongside questions about their implementation and the limited force of ethical commitments. Mittelstadt argues that agreement

on principles does not supply the institutional conditions needed to put them into practice. These are arguments about selected texts and governance arrangements, not proof that all organizations behave alike. [1] [2] [3]

Two further works sharpen the question. Ananny and Crawford challenge the assumption that making a system visible is sufficient for understanding or governing it. Raji and colleagues propose an internal auditing framework that organizes documentation across an AI system's development lifecycle. Together, these works motivate an examination of the connection between information and action. The framework developed below is this paper's proposed synthesis. It does not reproduce either source's framework, claim a newly discovered research gap, or describe an evaluated intervention. [4] [5]

Source selection and conceptual approach

The corpus consists of five deliberately selected works associated with publications from 2018 to 2020. Three examine AI ethics guidelines or principled governance, one critically examines transparency, and one develops an internal auditing framework. They were chosen to juxtapose normative commitments, the limits of visibility, and organizational documentation. This is a selective conceptual synthesis of foundational literature. It is not a systematic review, an account of the latest research, or a representative sample of institutions. The selection supports comparison among arguments; it cannot establish how frequently those arguments hold in practice.

Accessible full texts were retrieved on October 21, 2026. The inspected copies are identified individually in the references and accompanying curation record: three versioned arXiv manuscripts, an author-linked copy of Ananny and Crawford's published article, and Raji and colleagues' arXiv manuscript. Bibliographic publication dates are distinguished from manuscript revision dates. Relevant passages were checked against those copies. No new dataset, interviews, experiments, or organizational observations were produced. No database-wide search count, protocol registration, or independent human review is claimed. The public sample was prepared with AI assistance outside the product's research-job pipeline.

The analysis proceeds by defining constructs, proposing relationships among them, and examining counterarguments. Definitions introduced here are working definitions for this argument rather than claims that the cited authors share a single terminology. Source-based statements describe a work's argument or proposed method. Propositions and the organizational illustration express the present synthesis. Keeping those categories separate is necessary because a defensible interpretation of earlier scholarship is not equivalent to a demonstrated effect. The framework should therefore be assessed first for coherence, useful distinctions, and explicit boundaries. Any claim of effectiveness would require further evidence.

Construct definitions

For this paper, transparency means an arrangement that makes information about a specified system or decision available to a specified audience. The definition requires an object, an audience, and an access condition. An object could be the intended use of a model, a reason for a particular output, or the basis for approving deployment. The audience might be a person affected by a decision, an internal reviewer, or another oversight body. These possibilities should not be treated as interchangeable. A disclosure designed for one audience need not answer the questions of another, even when both can access the same document.

Interpretability is used here in a practical, audience-relative sense: whether the recipient can make relevant sense of the disclosed information for the question at hand. This definition does not equate interpretability with a particular technical explanation method. A document might be readable yet omit the assumptions necessary to challenge its conclusion. Conversely, a technically detailed account could be useful to a specialist while requiring mediation for other audiences. The analytical question is whether the information permits the intended recipient to identify and assess a reason for action. This is a proposed criterion, not a measured property of any system discussed in the sources.

Accountability is defined here as an arrangement in which an identifiable actor must explain a decision to an audience entitled to question it, with a specified process for responding.

Answerability concerns the obligation to give reasons. Corrective capacity concerns the ability and authority to change a decision or address a problem. The two are separated because a person might provide an explanation without being able to authorize a remedy. The framework does not assume that every challenge warrants a change. It requires a way to consider the challenge and communicate an accountable response, including a reasoned decision to retain the original action.

Relationships between the constructs

The proposed relationship begins with a decision rather than a document. A reviewer first identifies what is being decided and whose interests could be relevant. The next questions concern what information can be inspected, which audience can interpret it, and who is required to respond. This sequence is a heuristic for examining an arrangement. It should not be read as a universal chronological workflow. In an actual setting, disclosure, questioning, and revision could occur repeatedly or in parallel. The important feature is that an intelligible connection can be traced among them.

The framework treats transparency as a possible input to accountability rather than its equivalent. Information may enable a challenge, but an enabling condition is different from an obligation to hear that challenge. Likewise, assigning a responsible person does not show that an audience can obtain the information needed to question the decision. These observations follow from the

working definitions. Their practical value lies in preventing one feature from standing in for an entire governance arrangement. An assessment should identify which link is present, which is missing, and what evidence would be needed to judge its operation.

Documentation provides a proposed means of maintaining those connections. Raji and colleagues describe an internal audit process with documentary artifacts across development stages. Their proposal supports treating records as objects that can connect decisions across a lifecycle. It does not establish that documentation necessarily creates independent scrutiny or effective remedies. The present synthesis adds a narrower analytical requirement: a record should identify the decision it informs, the audience it serves, and the response process it enables. That requirement is an interpretation developed here, not an attributed result from the auditing paper. [5]

Proposition one Audience fit

The first proposition is that disclosure supports answerability more plausibly when its content and form are matched to an audience's task. Audience fit requires specifying the question the recipient is expected to examine. If the task is to question a deployment boundary, an account of intended uses and excluded uses may be more pertinent than an isolated description of a model's architecture. This is a conditional proposal about relevance. It does not imply that architectural information is generally unhelpful, or that a single explanation can satisfy all interested audiences. The appropriate content depends on the decision being challenged.

One way to examine the proposition would be to ask whether a recipient can connect a stated concern to a specific feature of the disclosed account. A successful connection would show that the account affords a route for questioning; it would not, on its own, prove that the resulting judgment is correct. The distinction matters because accessibility, comprehension, and justified evaluation are different achievements. Future work could investigate them separately. This paper does not conduct that investigation or recommend a universal comprehension threshold. It offers audience fit as a question that a more detailed assessment could operationalize in context.

Proposition two Response authority

The second proposition is that transparency is more closely connected to accountability when the recipient can direct a challenge to an actor with a defined obligation and authority to respond. An organization could name a contact while leaving that contact unable to request changes. It could also assign broad responsibility without defining how a concern reaches the responsible person. Both hypothetical arrangements would leave the relationship between information and action incomplete under this paper's definition. The proposition calls for an explicit connection between a challenge, a decision owner, and the range of responses available to that owner.

Response authority should be distinguished from automatic agreement with the challenger. A process might permit clarification, further investigation, a revised deployment condition, or a justified refusal to change the decision. What makes those responses analytically relevant is their relation to an obligation to consider the issue. The framework also leaves room for escalation when a decision owner cannot resolve it. These are design questions derived from the proposed account of accountability. They are not findings about existing complaint systems, evidence of legal compliance, or proof that any particular organizational structure will produce fair outcomes.

Proposition three Continuity across decisions

The third proposition is that records support continuing accountability when they preserve the relationship between an earlier reason, a later decision, and the actor responsible for reconsideration. A disclosure could become difficult to use if the system it describes changes while the document's scope remains unclear. Within the framework, the relevant question is whether a reader can determine which decision and system state the record concerns. Version identifiers, dates, and documented reasons are possible ways to make that relationship inspectable. Their inclusion is proposed as a traceability practice, not as a sufficient test of responsible behavior.

Continuity also requires a rule for reopening questions. For example, a change in the intended use of a hypothetical system might require reconsidering the reasons previously offered for deployment. The analytical demand is to state which changes trigger review and who decides whether the earlier account still applies. The proposition does not require permanent disclosure of every internal detail or unlimited record retention. It requires a defensible explanation of how relevant decisions remain connected over time. The costs, privacy implications, and practical effectiveness of particular implementations remain outside the evidence supplied by this paper.

Hypothetical organizational illustration

Consider a fictional organization deciding whether to deploy a model that prioritizes internal support requests. It prepares a short intended-use statement and a technical assessment. Under the first proposition, the organization would identify the audience for each document and the questions that audience should be able to ask. Staff who could be affected might need an explanation of the prioritization criteria and a route to raise concerns. A technical reviewer might instead examine assumptions in the assessment. This example supplies no empirical claim about support systems, their users, or the adequacy of particular criteria.

Under the second proposition, a named decision owner would receive questions and explain what could change following review. Under the third, the decision record would connect the assessed version, reasons for deployment, and conditions that require reconsideration. Imagine that a new use is proposed after approval. The record would help identify which earlier reasons require

review rather than treating the existence of a document as continuing authorization. The example illustrates the framework's vocabulary. It does not demonstrate improved fairness, accuracy, trust, or institutional performance, and it should not be interpreted as implementation advice for a regulated application.

Counterarguments and boundary conditions

A useful distinction within the framework is the difference between a missing connection and a disputed reason. If an audience can identify a decision owner and obtain a response, the basic relationship may exist even when the audience rejects the reasons offered. That disagreement should remain open to substantive criticism. By contrast, if no actor can be required to answer, the difficulty concerns the accountability arrangement itself. Separating these cases helps locate the work an assessment must do. One case calls for examining the adequacy of reasons, while the other calls for examining the conditions under which reasons become available for challenge.

A first objection is that the framework makes transparency depend on organizational conditions beyond disclosure. That objection is reasonable if transparency is defined narrowly as access to information. The present account preserves that narrow distinction while asking when access contributes to accountability. It does not deny the independent value that disclosure might have for inquiry or public discussion. Instead, it limits the inference that can be drawn from disclosure alone. An institution could meet a transparency commitment while leaving questions about response authority unanswered. The framework is intended to make those separate judgments explicit.

A second objection concerns internal documentation. Records might support internal coordination without giving external audiences influence over decisions. Ananny and Crawford's critique makes the relationship between visibility, understanding, and governance a relevant concern, while Raji and colleagues focus on a proposed internal audit framework. These different emphases should remain visible. The present synthesis does not resolve the problem of institutional independence. It asks that the identity and authority of each audience be stated so that internal review is not silently treated as external oversight. Whether a specific oversight arrangement is sufficiently independent requires a separate argument. [4] [5]

A third objection is that the proposed connections could impose burdens disproportionate to a decision's consequences. The framework offers no numerical rule for proportionality. It instead requires the scope of the assessment to be stated. A bounded explanation of a low-consequence internal choice and a sustained challenge to a consequential deployment need not involve identical procedures. However, proportionality should itself be explained rather than assumed. A useful next step would be to compare competing accounts of what information, audiences, and response obligations a specified context warrants. No such comparative evaluation is reported here.

Limitations and implications

The corpus is small, purposive, and confined to selected foundational works. It cannot support conclusions about the full literature, current governance practice, or the distribution of views across disciplines and regions. Accessible manuscripts also differ from final publications in ways recorded with the references. Statements are restricted to the consulted versions. The synthesis gives little attention to particular legal systems, domain-specific institutions, or the perspectives of people affected by an actual deployment. These exclusions limit its use as a practical governance model and prevent claims that its propositions are generally valid.

The most defensible implication is a set of questions for conceptual appraisal. What decision is under examination? What information reaches which audience? What makes it usable for a challenge? Who must respond, and what can that actor change? What preserves the connection when the decision or system changes? Asking these questions may clarify an argument, but clarity is not an effectiveness result. Further scholarship could challenge the definitions, identify excluded relationships, or test the propositions in a specified setting. Such work would need its own methods, evidence, and ethical consideration.

Conclusion

This paper proposes that transparency contributes to accountability through relationships among accessible information, an audience capable of questioning it, and an actor obliged to respond. Audience fit, response authority, and continuity across decisions express different parts of that relationship. The distinction helps specify what a disclosure arrangement establishes and what remains to be justified. It also leaves disagreement about institutions, purposes, and remedies open to examination. As a selective conceptual synthesis, the paper offers working definitions and conditional propositions rather than empirical conclusions. Its value depends on whether those distinctions support more precise reasoning about a particular decision, not on an assumption that visibility guarantees responsible action.

References

Publication details identify the associated work. Version notes identify the full texts inspected.

[1] Anna Jobin, Marcello Ienca, Effy Vayena (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1, 389–399. <https://doi.org/10.1038/s42256-019-0088-2>

arXiv:1906.11668v1, submitted June 24, 2019. Inspected preprint title: Artificial Intelligence: the global landscape of ethics guidelines.

[Read inspected version](#)

[2] Thilo Hagendorff (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30(1), 99–120. <https://doi.org/10.1007/s11023-020-09517-8>

arXiv:1903.03425v2, revised October 11, 2019; preprint of the 2020 article. It is not identical to the corrected journal text; no guideline counts are used in this sample.

[Read inspected version](#)

[3] Brent Mittelstadt (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1, 501–507. <https://doi.org/10.1038/s42256-019-0114-4>

arXiv:1906.06668v2, revised February 18, 2020; author manuscript associated with the 2019 publication.

[Read inspected version](#)

[4] Mike Ananny, Kate Crawford (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>

Author-linked copy of the paginated 2018 journal article; online-first publication December 13, 2016. Retrieved through the public author-linked download.

[Read inspected version](#)

[5] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, Parker Barnes (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. <https://doi.org/10.1145/3351095.3372873>

arXiv:2001.00973v1, submitted January 3, 2020; conference manuscript.

[Read inspected version](#)

Preparation record

Source assertions were compared with identified passages using AI-assisted editorial review. The accompanying curation record contains source versions, access dates, paragraph classifications, support checks, and limitations. No independent human review or product approval is claimed.

Application safeguards are described on the accompanying website’s Methods & quality page.