

From Principles to Practice: Measurement as the Bridge Between Ethical AI and Responsible AI

Syed Mohammad Shayaan Hasnain
shayanhasnain101@gmail.com

ABSTRACT

The rapid evolution of Artificial Intelligence (AI) has increased the need for organizations to translate ethical principles into actionable and governable practices. While ethical AI establishes normative principles such as fairness, transparency, and explainability, responsible AI requires mechanisms to operationalize and govern them.

This review examines measurement as the missing connective layer between ethical AI and responsible AI and investigates the resulting standardization gap.

A purposive narrative review of relevant literature was conducted to examine ethical constructs, measurement approaches, and standardization challenges.

The synthesis indicates that fairness is extensively measured through multiple and non-comparable metrics, while transparency and explainability remain comparatively insufficiently measured and lack widely validated measures. The paper argues that effective responsible AI requires a measurement approach that connects clearly defined ethical constructs with valid and reliable measures, stakeholder input, and explicit governance of competing objectives.

The paper contributes by reframing the principles-to-practice gap as a measurement problem and proposing a bridge between ethical AI and responsible AI.

Keywords: Ethical AI, Responsible AI, Fairness, Transparency, Explainability, AI governance, Standardization,

INTRODUCTION

Johnson et al. (2025) posited, AI is not only limited to specialized research, but also into a pervasive sociotechnical infrastructure. Moreover, with the invention of agentic AI it's steering decisions in finance, education, aerospace, retail stores, healthcare, transportation and the public sector. Notably, UAE has

Sep 2026
Vol 11. No 3.

already set a goal in 2025 to automate around 50% of the governmental services and operations with agentic AI by 2027 (UAE Government, 2025) and Singapore debuts into the world's first governance framework with agentic AI (IMDA, 2026).

This rapid rise in AI Adoption and governmental initiatives has been accompanied by an extensive normative literature that follows the value that AI systems ought to respect. Jobin et al. (2019) argued about the global landscape of AI ethics guidelines, with principles i.e. transparency, justice and fairness across documents are crucial to be produced by governments, companies, and professional bodies.

The discipline has accordingly bifurcated into two distinct area. The first, frequently labelled ethical AI, is normative, it asks what AI systems should do and which values should compel their design and deployment (Floridi & Cowls, 2019). The second, is responsible AI, which is operational and answers, how organizations can ensure that those values are realized in practice, and who is accountable when they are not. Hence, responsible AI is the machinery built to honor the commitments that ethical AI articulates (Mittelstadt, 2019). Yet many studies are evident that the computational systems and processes which are meant to make AI ethical usually don't work into that direction (Morley et al., 2021).

Mittelstadt (2019) argued that principles alone cannot guarantee ethical AI. Additionally, Hagendorff (2020) demonstrated, although the guidelines are rapidly increasing across the field are vague, mutually redundant, and disconnected from any mechanism of enforcement. The differentiation between stated principle and realized practice has been indicated in this paper as a principle to practice gap (Mittelstadt, 2019; Morley et al., 2021; Rakova et al., 2021 and Veale, Van Kleek & Binns, 2018).

Research Gap and Objective

This paper advances a specific diagnosis of why that gap persists. The argument is that the connective tissue between ethical AI and responsible AI is measurement, and that this tissue is underdeveloped and unstandardized (Mittelstadt, 2019; Morley et al., 2021). Responsible AI is a governance project, and governance requires the capacity to assess whether a system meets a stated expectation; that capacity, in turn, requires valid and comparable measurement of the relevant property (Rakova et al., 2021; Veale et al., 2018). Where a principle cannot be measured consistently, responsible AI is asked to govern something it cannot observe, and the result is either superficial compliance or the displacement of ethical commitments by whatever happens to be easy to quantify.

The paper further argues that the measurement problem is not uniform across principles. Instead, it exhibits structural asymmetry. For fairness, the field suffers from an excess measurement; a large and growing inventory of formal metrics, many of which are mutually incompatible (Mittelstadt et al., 2019 ; Doshi-Velez and Kim, 2017). For transparency and explainability, the field suffers from a deficit: the absence of agreed, standardized performance measures at all (Morley et al., 2021; Rakova et al., 2021). In both cases, and across the principles taken together, there is no accepted standard for adjudicating the

trade-offs that arise when several criteria must be satisfied simultaneously. The paper terms this twofold condition the standardization gap.

This research paper contribution is threefold. First, the paper offers a conceptual account of measurement as the connective layer between ethical AI and responsible AI, clarifying why operationalization is impossible without it. Second, it indicates standardization gap as a structural, which shows the way to solve the problem (ReyesCruz & GarcíaSánchez, 2025). Third, it proposes a research and governance agenda for standardization i.e. three separable tasks. Accordingly, the findings revolve around three research questions:

Research Question 1: What relationship links ethical AI to the responsible AI?

Research Question 2: What is AI-ethics measurement, and reason for standardization in this approach not yet emerged?

Research Question 3: How can standardization of ethics measurement bridge the gap between principle and practice?

To answer the above questions, this study adopts a purposive narrative literature review. The selected literature is then analyzed thematically. The analysis then compares the measurement landscape for fairness, transparency and explainability and uses the identified gaps to develop the proposed conceptual framework.

METHODS

This section outlines the methodology of literature review i.e., the approach by which the sources were selected for the paper, literature identification, inclusion & exclusion along with the selection and analytical process as follows.

Review Approach

The aim of this paper is conceptual rather than empirical. Therefore, the research paper adopts a purposive narrative literature review approach, to examine the conceptual relationship between ethical AI and responsible AI, and to synthesize relevant literature and critically examine principles to practice translation (Jaakkola, E. ,2020). Unlike a systematic review, which uses standardized search and screening procedures to minimize selection bias in synthesizing an evidence base, the present purposive approach selects conceptual depth capable of informing the development of the research based conceptual argument over comprehensive coverage (Jacobs, A. Z., & Wallach, H., 2021).

Literature Identification and Search Scope

Sep 2026
Vol 11. No 3.

The literature search focused on research published between 2022 and 2026 in order to capture the rapidly developing evidence based on responsible AI, AI ethics measurement, evaluation, and operationalization. Earlier foundational studies were also included selectively where provided, theoretical, methodological or conceptual contribution for this research. These include foundational work on AI ethics principles, the principles-to-practice gap, fairness measurement, sociotechnical abstraction, and measurement theory (Jobin et al., 2019; Mittelstadt, 2019; Selbst et al., 2019; Jacobs & Wallach, 2021). Their inclusion was therefore based on conceptual significance rather than publication recency.

Attention was given to peer-reviewed research indexed in Scopus, and the ACM Digital Library and Artificial Intelligence review etc. The search was organized around following concepts; “AI ethics,” “ethical AI,” “responsible AI,” “fairness metrics,” “explainability,” “interpretability,” “measurement,” and “construct validity.”

Inclusion and Exclusion Criteria

Literature was selected using the following inclusion criteria:

1. The publication directly addressed ethical AI, responsible AI, trustworthy AI, AI governance, AI ethics, or the operationalization and evaluation of ethical principles.
2. The study contributed to at least one of the analytical dimensions of this paper: measurement, operationalization, fairness, transparency, explainability, governance, standardization, or ethical trade-offs.
3. The publication was peer-reviewed where it was used as core scholarly evidence.
4. Preference was given to research published between 2022 and 2026 to reflect the current development of responsible AI research.
5. Highly cited or conceptually foundational earlier studies were retained where their theoretical contribution was necessary to establish the conceptual basis of the study.
6. Systematic reviews, structured literature reviews, surveys, measurement studies, and evaluation frameworks were given particular attention because they provide evidence concerning the organization and maturity of the existing measurement landscape.

Excluded studies were, primarily unrelated to ethical or responsible AI, focused exclusively on technical model performance without relevance to ethical evaluation or governance, duplicated arguments already represented by stronger synthesis studies, or did not provide sufficient conceptual or empirical relevance to the research questions. Policy, regulatory, and standards documents were considered separately from the peer-reviewed literature because they provide institutional and governance context rather than equivalent scholarly evidence.

Selection and Analytical Process

The literature was analyzed through a thematic synthesis rather than through statistical meta-analysis.

First, selected studies were examined for their treatment of the relationship between normative ethical principles and their practical implementation.

Second, relevant findings were organized according to the major analytical constructs emerging from the research questions: (1) Ethical AI as a normative layer, (2) Responsible AI as an operational layer, (3) the principle-to-practice gap, (4) measurement and operationalization, (5) standardization of ethical measures, and (6) trade-offs between competing principles.

Third, literature was compared across the three principles that are central to this study; fairness, transparency, and explainability. The analysis specifically examined whether each principle has clearly defined constructs, established operational measures, validated evaluation approaches, and mechanisms for resolving conflicts with other principles. Finally, the findings were synthesized to identify recurring structural patterns rather than merely summarize individual studies.

LITERATURE REVIEW

Ethical AI as the Normative Layer

Artificial intelligence is transforming and revolutionizing science, societies and human intelligence by bringing in automated ways to perform tasks and take decisions. While ML, artificial neural networks, deep learning and agentic AI are reinventing the automation processes verticals such as transportation, healthcare etc. are extensively adopting AI for efficiency and efficacy. Hence, the impact of such disruptive technologies can be observed into various societal domains. Despite, being obvious surge in business growth there are ample discourse about the principles and values, which are guiding mechanisms of such developments (Jobin et al., 2019).

Moreover, Jobin et al. (2019) critiqued ethical AI as a normative layer for specifying the values that AI systems to follow. In turn, its outputs are principles, declarations, and codes of conduct, and its intellectual base runs through applied ethics and political philosophy. Accordingly, Floridi and Cows (2019), proposed a unified framework of five core principles in bioethics i.e. beneficence, non-maleficence, autonomy, justice and explicability. The uniqueness of such framework lies in their generality, across applications and jurisdictions. However, the generality is also being critiqued by many research as a operational weakness, due to its high level of abstraction which by itself doesn't specify what compliance would look like (Awad et al., 2018).

Sep 2026

Vol 11. No 3.

Skeptics have also questioned the credibility of principle-based ethics which inherit limitations from medicine that do not sit well with AI Systems. Mittelstadt (2019) observed that medicine benefits from common aims, professional norms, robust accountability mechanisms, and a long history of translating principles into practice, are missing in AI development. The implication is that ethical AI, however well specified at the level of value, cannot comply with its own commitments. It requires a second layer that operationalizes principle into practice.

Responsible AI as the Operational Layer

Morley et al. (2021) reported that responsible AI is implicated towards governance frameworks, audits impact assessment, maturity models and assignments of responsibilities by being an integral part of ethical AI, playing a role as a controller and forming accountability through organizational ethical commitments, across the AI landscape.

Therefore, framing of ethics must be operationalized rather than merely declared as a governing principle to deliver ethics as a service, embedding ethical review into the development workflow (Reyes-Cruz et al., 2025). Hence, auditing has become a central instrument of this layer which possesses structured, multi-layered approaches to auditing large language models. Thus, it illustrates the attempt to convert normative expectations into iterative assessment procedures as second layer of operationalization (Mokander et al., 2024).

Notably, some authors treat responsible AI as the broader umbrella, encompassing ethical, legal, robust, and sustainable dimensions, with ethics as one constituent. Others treat ethical AI as the broader normative horizon and responsible AI as its implementation or operational layer (Reyes-Cruz et al., 2025; Morley et al., 2021; Jobin et al., 2019).

This paper adopts the convention which states that ethical AI is the normative layer that specifies what ought to be achieved, and responsible AI is the operational layer that specifies how it is achieved and by whom (Gyevnár, B., & Kasirzadeh, A., 2025). Therefore, this paper discussed two distinct layers and the how the connection between them must be established.

Principles-to-Practice Gap

The persistent challenge in AI ethics is translating normative commitments into daily organizational practices. Hence, the principles to practice gap evaluated in this paper explains the tendency of institutions to adopt ethical frameworks, while failing to adopt this commitment into the design, development and deployment of Computational systems (Jobin et al., 2019).

Hagendorff (2020) critiques that ethical declarations often remain disconnected from operational decision-making, giving rise to what has been described as ethics washing, where the appearance of ethical responsibility replaces meaningful implementation.

Further evidence of this disconnect is provided by recent synthesis research examining the operationalization of AI ethics in computing. Rismani et al. (2025) reviewed the measures used by researchers to assess ethical principles and found that evaluation approaches are fragmented, narrowly focused, and unevenly into existing literatures.

This lack of consistency limits comparability, and weakens critical evaluation, toward accountable AI governance. Hence, the implementation gap is persistence, despite the usage of ethical guidelines from past decades. As a result, it, indicates that the central challenge is no longer identifying ethical principles but establishing rigorous, systematic methods for verifying whether those principles have been effectively translated into practice (Mittelstadt, 2019).

CONCEPTUAL MODEL AND PROPOSITIONS

Constructs and Operationalization

Jacobs and Wallach (2021) defined constructs as unobservable theoretical entities and must be inferred from observable properties via a measurement model. Therefore, constructs require explicit definition before measurement because different operationalizations may represent different interpretations of the same normative concept. This distinction is particularly important for fairness, where alternative metrics may encode different underlying conceptions of what fairness requires (Jacobs & Wallach, 2021).

Adcock and Collier (2001) defined operationalization is the process of translating an abstract construct into observable indicators, variables, procedures, or assessment rules. Operationalization therefore precedes measurement: the construct establishes what is intended to be assessed, while operationalization establishes how evidence of that construct will be obtained. Responsible AI aspires to comparable governance, but the properties it must govern, such as fairness, transparency, and explainability are not directly observable in the way that a physical quantity is. They are abstract constructs that must be operationalized, that is, connected to observable indicators through an explicit procedure (Rismani et al., 2025; Gyevnár & Kasirzadeh, 2026). Without such a procedure, an expectation cannot be checked, an audit cannot reach a verdict, and accountability cannot attach to a determinate finding. Hence, measurement is necessary component for refinement of responsible AI, and it converts a normative commitment into an assessable claim (Figure 1).

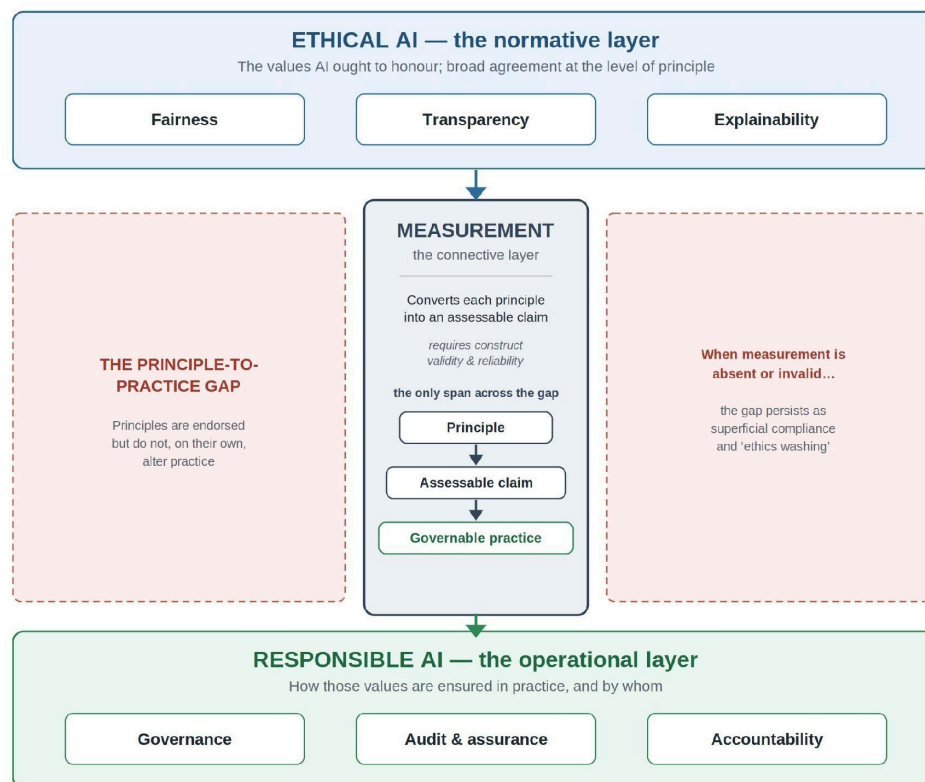


Figure 1: The two-layer structure of ethical AI and responsible AI

Figure 1 shows that ethical AI specifies principles on which there is broad agreement (top); responsible AI must render those principles into governable practice (bottom). The two layers are separated by the principle-to-practice gap. Measurement is the connective layer that spans the gap by converting each principle into an assessable claim; where measurement is absent or invalid, the gap persists as superficial compliance. The schematic synthesizes the principle-based critiques of Mittelstadt (2019) and Hagendorff (2020), clarifies why principle-based ethics stalls implementation. The move from ethical AI to responsible AI is precisely the move from naming a value to measuring its realization, and it is at this junction that the field's difficulties concentrate.

Measurement, Validity and Reliability

Measurement is conventionally defined as the assignment of values to attributes according to an explicit rule (Stevens, 1946). Jacobs and Wallach (2021) reframed this as measurement modeling: because constructs are unobservable, values are not assigned directly but inferred from observable properties through a model whose assumptions link construct to indicator.

A measurement is therefore interpretable only when four elements are made explicit: the construct being assessed, the operationalization adopted, the procedure applied, and the context in which the resulting scores are read. Adcock and Collier (2001) treat this last element as substantive rather than incidental, arguing that measurement claims are contextually specific and that the interpretation of scores in relation to a systematized concept constitutes a falsifiable claim requiring supporting evidence, not a definitional guarantee. Measurement is thus not quantification alone, but the disciplined representation of a defined construct through a procedure that is explicit enough to be evaluated and, where the assumptions fail, contested (SuárezÁlvarez et al., 2026). Hence, the existence of a measurement does not establish that the measurement is fit for governance. Two properties are particularly important: validity and reliability.

Validity concerns whether the evidence produced by a measurement procedure adequately represents the construct that the measure is intended to assess (Adcock & Collier, 2001). Jacobs and Wallach (2021) demonstrated the disagreement among fairness metrics can reflect unresolved disagreement about the underlying construct rather than merely disagreement about numerical thresholds.

Validity should therefore be treated as a property requiring justification rather than as an automatic consequence of using a formal metric (Salaudeen et al., 2025). In the context of ethical AI, validity requires consideration of whether the construct has been adequately defined, whether the operationalization corresponds to that definition, and whether the resulting evidence supports the intended interpretation. Applied to responsible AI, construct validity therefore asks whether a selected indicator captures the relevant ethical property rather than a simplified or unintended substitute for that property.

On the other hand, reliability concerns the consistency and stability of a measurement procedure under comparable conditions (Adcock & Collier, 2001). A reliable procedure should not produce substantially different results merely because a measurement is repeated or because different assessors apply the same procedure. Jacobs and Wallach (2021) emphasize both validity and reliability to be considered for defensible measurement in computational systems, noting that both must be assessed rather than assumed. Reliability and validity are nonetheless distinct properties. A measure may be reliable but invalid, repeatedly producing the same result while consistently capturing the wrong property (Adcock & Collier, 2001).

The distinction can therefore be represented as follows the sequence: construct definition, operationalization, measurement, validity and reliability assessment, interpretation, governance decision

This sequence strengthens the proposed measurement bridge (Figure 1) because it prevents responsible AI governance from treating any numerical result as automatically meaningful. Measurement becomes a governance instrument only when the construct is explicitly defined, the operationalization is justified, the resulting measurement is sufficiently valid and reliable for its intended purpose, and the conditions of interpretation are documented (Blodgett et al. 2021; Rismani et al. 2025).

THE STANDARDIZATION GAP

Standardization refers to the development and adoption of agreed definitions, procedures, measurement rules, reporting requirements, and decision criteria that enable an ethical property to be assessed consistently and meaningfully compared across relevant systems and contexts (Timmermans, S., & Epstein, S., 2010). Standardization does not necessarily require one universal metric for every AI system. Given the sociotechnical and contextual nature of ethical constructs, a more appropriate objective is methodological standardization: agreement on how constructs are defined, how measures are selected, how validity and reliability are assessed, how measurement conditions are reported, and how differences or trade-offs are justified (Adcock and Collier, 2001).

Hence, this section argues the present state of non-uniformity and asymmetrical edges of ethical AI i.e. fairness, transparency and explainability. These have been evaluated based on present literature and trade-offs are being examined.

Fairness

Fairness is the principle for which formal measurement is most advanced, and keeping the source of its difficulty. A large inventory of statistical fairness criteria now exists, including demographic parity, equality of opportunity, and predictive parity, each formalizing a distinct intuition about what equal treatment requires (Hardt et al. 2016; Mehrabi et al. 2021). The profusion would be benign if the criteria agreed, but it's not the case. Results indicate that several intuitively compelling fairness criteria cannot be satisfied simultaneously except in degenerate cases. When base rates differ across groups, calibration and equalized error rates are mutually incompatible (Kleinberg et al. 2017; Chouldechova, 2017).

The consequence of governance is that an organization can demonstrate fairness on one metric while a critic demonstrates unfairness on another metric. Hence, there is no standard available to resolve between them. Notably, fairness is, in this sense, over-measured i.e. the abundance of metrics, absent any principled rule for selecting among them, produces not clarity but the possibility of selective and self-serving reporting. The condition described earlier as conflation of construct and operationalization (Jacobs and Wallach 2021) is thereby exhibited at scale.

Transparency

Transparency refers to the availability and accessibility of relevant information concerning an AI system, including its purpose, development, data, operation, limitations, governance arrangements, and relevant decision processes. Transparency is consequently broader than the explanation of an individual model output (Jobin et al., 2019).

A transparent system may provide information about what the system is intended to do, what data and processes support it, who is responsible for it, what limitations apply, and how its performance and risks are monitored. Transparency is therefore concerned substantially with the availability and accessibility of information required by relevant stakeholders (Mitchell et al., 2019).

Explainability

Explainability refers to the extent to which the behavior or outputs of an AI system can be meaningfully explained to an identified audience for a specified purpose. Explainability is therefore concerned with intelligibility and understanding rather than merely information availability (Barredo Arrieta et al., 2020). Transparency and explainability are related but distinct. An organization may provide extensive technical documentation and therefore demonstrate a degree of transparency while still failing to provide explanations that are understandable or useful to affected users (Mitchell et al., 2019; Lipton, 2018).

Critics argued that the field had not established a rigorous basis for evaluating interpretability at all, leaving the notion under-specified and its assessment largely informal (Doshi-Velez and Kim, 2017). The difficulty is compounded by a normative dispute about whether post hoc explanation of opaque models is even the right goal, given arguments that high-stakes decisions should rely on inherently interpretable models rather than explanations layered onto black boxes (Rudin, 2019).

A recent systematic review of more than three hundred explainability studies reinforces this diagnosis, that evaluation practices remain largely qualitative and inconsistent and that no shared quantitative standard yet exists for judging the adequacy of an explanation (Nauta et al., 2023). In addition, a survey of evaluation criteria for trustworthy AI note that, while quantifiable metrics can be computed for properties such as fairness and accuracy, no comparable performance metrics have been established for many other principles, so their assessment defaults to manual or procedural review rather than measurement (Rismani et al., 2025). As a result, transparency and explainability are insufficiently standardized measurements. In a precise sense, the measurement layer of responsible AI would be needed for governance.

Trade-off Management

Kleinberg et al. (2017) identified that, ethical principles are not independent; satisfying one can compromise another. When fairness, accuracy, transparency, and privacy are examined in a tension, it has shown impossibility results for fairness, so that improvement on one dimension may degrade to another. Hence, a standardized measurement-based framework would be needed to fix metrics and threshold for each principle in isolation. The specification of how trade-offs are to be represented, weighted and justified should also be established.

Due to absence of such standards to justify each metric, resolution of trade-offs is made implicitly and opaquely within AI system development. Hence, the standardization gap, has three concerns: too many

fairness metrics, too few transparency metrics, and no accepted account of the trade-offs that bind them together (Nauta et al., 2023).

WHY STANDARDIZATION HAS NOT EMERGED

Sociotechnical Characteristics of Ethics

A first obstacle is that the constructs at issue are contested and unavoidably sociotechnical. Concepts such as fairness, transparency, and explainability do not have universally fixed meanings. Rather, they are normative constructs whose interpretation depends on the social, cultural, and institutional contexts in which they are deployed (Johnson et al., 2025).

As Selbst et al. (2019) argue, translating these concepts into formal technical metrics can produce abstraction pitfall, in which the broader societal context that gives them significance is removed. As a result, computational measures may assess simplified replacement instead of the underlying ethical concerns they are meant to capture. If the construct legitimately varies with context, then a single universal metric is not merely difficult to construct but conceptually inappropriate, that's, why metric proliferation is critical, rather than convergence.

Contextual and Cultural Variability

A second obstacle is variability across domains and cultures. The interpretation of an ethical principle, and the indicators that would count as evidence of its realization, differ across application settings and national contexts, so that a measure validated in one setting may be invalid in another. The instruments for ethics-related constructs connect both the feasibility and the fragility of measurement. While, validated frameworks can be constructed, they can also calibrated to particular populations and purposes and do not transfer automatically (Yang et al., 2025). Standardization that ignored this variability would risk imposing an apparent uniformity. In contrast, standardization that fully accommodated it would risk dissolving into incomparable local measures. Navigating between these failure modes is an actual and unresolved design problem (Selbst et al., 2019).

Institutional and Regulatory Fragmentation

A third obstacle indicates institutional and regulatory fragmentation. Standardization is not only related to a scientific achievement but its a coordination problem requiring authoritative bodies, shared procedures, and incentives for adoption (Maslej et al., 2025).

The present landscape is divided across competing governance frameworks such as the NIST AI Risk Management Framework (National Institute of Standards and Technology. 2023), process-oriented management-system standards such as ISO/IEC 42001 (International Organization for Standardization

2023), and emerging regulation such as the European Union's Artificial Intelligence Act (European Union, 2024), with no single governing authority.

Risk-based AI regulations need developers of high-risk systems to provide measurable evidence for satisfy their model for ethical and performance requirements i.e. equitable performance across demographic groups. However, current compliance processes rely predominantly on procedural governance frameworks and management standards instead of validated and standardized performance measures. This creates a disconnect between regulatory expectations and the measurement capabilities available to organizations (Rismani et al., 2025). At the same time, comprehensive reviews indicate that although new evaluation benchmarks are emerging, universally accepted standards for assessing responsible AI remain limited (Maslej et al., 2025). Consequently, isolated measurement approaches, regardless of their individual quality, are unlikely to support consistent and comparable AI governance without broader institutional coordination and standardization.

RESULTS: A PROPOSED FRAMEWORK

The diagnosis above implies that standardization should not be pursued as a single undertaking of fixing metrics. The difficulties differ by principle and operate at distinct levels. This section proposes disaggregating standardization into three separable but linked tasks, which together constitute the bridge between ethical and responsible AI (Figure 2).

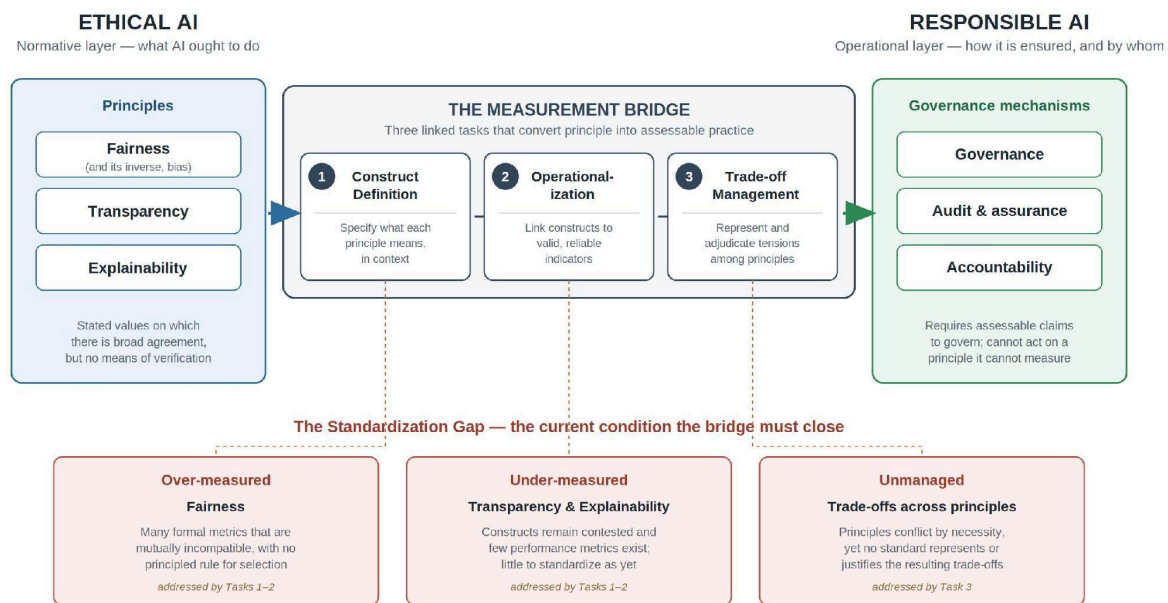


Figure 2: The measurement bridge between ethical AI and responsible AI.

The first task is construct definition:

Stating, for each principle and for a specified context, what property is to be assessed, in terms precise enough to be operationalized and explicit about the value commitments it encodes. The critiques of conflation and abstraction indicate that this task is logically prior to any choice of metric, since disagreement among metrics typically reflects unresolved disagreement at the level of the construct (Jacobs and Wallach, 2021; Selbst et al., 2019). For transparency and explainability, where constructs remain contested, this task is the binding constraint and must precede metric development (Lipton, 2018).

The second task is operationalization:

Connecting a defined construct to observable indicators through a procedure whose construct validity and reliability are themselves evaluated and reported, rather than assumed. This reframes the over-measurement of fairness as a problem of selection and justification under explicit criteria, and it directs effort for transparency toward building the interaction-based performance metrics that are currently absent.

The third task is trade-off management:

Specifying how the assessments of distinct principles are to be combined and how tensions among them are to be represented and adjudicated when they conflict. Because formal results guarantee that such conflicts are unavoidable (Kleinberg et al., 2017; Chouldechova, 2017), a standardization regime that addressed each principle in isolation would be incomplete by construction. Making trade-off resolution explicit and contestable is what would allow responsible AI to render accountable the value choices that are currently made silently within development.

The Framework in Practice: An Illustrative Worked Example

This section demonstrates how the proposed framework could operate in practice. Therefore, an illustrative hypothetical setting is used for an organization deploying AI system to screen job applicants, in which fairness, accuracy, privacy, and explainability are simultaneously at stake.

Table 1: The three stages of the proposed framework in practice

Stage	Key actors	Key activities	Expected output
1. Construct definition	Domain experts	Define ethical construct	Written construct specifications defining the selected construct, alternatives considered, contextual rationale, stakeholder concerns, and unresolved
	Affected-group representatives	Identify affected groups and potential harms	
	Legal/compliance staff	Consider alternative interpretations	

Stage	Key actors	Key activities	Expected output
	Ethics/AI-governance board	Document stakeholder Disagreement and uncertainty	uncertainties
2. Operationalization	Data scientists Measurement/validation specialists Domain experts Independent auditors Affected stakeholders where feasible	Select indicators and metrics Assess validity and reliability Evaluate data quality Establish thresholds Document measurement uncertainty and limitations	Measurement and validation specification containing selected measures, validity/reliability assessment, thresholds, assumptions, uncertainty, and limitations
3. Trade-off management	Accountable decision-owner Ethics/AI-governance board Compliance Privacy/risk representatives Relevant stakeholders	Compare competing objectives Assess consequences Consider stakeholder priorities Determine acceptable thresholds Select mitigation Document residual risk	Governance adjudication specifying the decision, trade-off rationale, accepted thresholds, mitigation measures, residual risks, approval, and monitoring requirements

Construct Definition

For instance, in recruitment, fairness can be defined by the organization, making sure that one protected group is not unfairly judged during shortlisting process. The meaning of fairness is decided explicitly, because fairness can mean different in different situations. Alternative interpretations, such as treating

every applicant equally based on individual merit or ensuring that different groups have similar error rates.

Stakeholders therefore have an important role in deciding what fairness should mean in specific scenario. In case there is a different opinion about the meaning of fairness, the organization should record the selected definition. Instead, it can give reason about its selection reason. This makes the value behind AI System clear and allows it to be reviewed or challenged later, instead of hiding that decision within the technical measurement process. (Jacobs & Wallach, 2021; Selbst et al., 2019).

Operationalization

At operationalization, the broad construct such as fairness is turned into practical way of measuring them. For fairness, organizations may need to choose among competing criteria, such as demographic parity and equalized odds. The choice should depend on the purpose and context of the system, not simply on which measure is easier to calculate (Hardt et al., 2016; Kleinberg et al., 2017).

Accuracy should be tested using data that properly represents the people and situations involved. Suitable measures should be used, including calibration when it is important for the decision being made. For explainability, it is not enough for an AI system to produce a technical explanation. The explanation should also be understandable to the people who use it, such as recruiters and, when appropriate, job applicants.

Uncertainty should also be reported when the available data allows it. This can include confidence or uncertainty intervals and clear information about the limitations of the data. These limitations may include missing or unreliable group information, small or unrepresentative samples, poor data quality, measurement errors, and assumptions made during the analysis.

Therefore, operationalization does not assume that one metric fully captures a concept such as fairness or accuracy. Instead, it makes clear what the chosen measure can show, what it cannot show, and how much confidence can be placed in the results.

Trade-off Management

For example, increasing a fairness threshold in AI system may reduce predictive accuracy. Research also demonstrate that several commonly desired fairness criteria cannot simultaneously be satisfied under certain conditions when group base rates differ (Kleinberg et al., 2017; Chouldechova, 2017). Similarly, providing richer explanations may reveal personal or sensitive information and therefore create tension with privacy. Conversely, limiting access to information about protected groups may help protect privacy but can make it harder to identify and monitor unfair outcomes.

Stakeholders should be involved when these trade-offs are considered. Their views can help determine whether a particular balance is reasonable, especially when the effects of a decision may place a greater burden on some groups than others.

The organization should record the final decision clearly. This record should state which objectives were given greater importance, the acceptable performance levels for the selected measures, the remaining risks, and the reasons behind the decision. The person responsible for the decision and the relevant governance body should formally approve it. This documentation can then be used for future audits, reviews, monitoring, and accountability.

The process is iterative rather than strictly linear. Where measurement evidence, stakeholder feedback, or governance review reveals that the construct or operationalization is inadequate, the organization may return to an earlier stage and revise the construct definition, measurement approach, or decision criteria.

DISCUSSION

Findings support the bifurcation between ethical AI and responsible AI as a normative and operational layers, conceptualizing the missing links i.e. as a measurement bridge between ethical AI and responsible AI. Therefore, successful justification through conceptualizing was given to answer three question i.e. What relationship links Ethical AI to Responsible AI? What is AI-ethics measurement, and the reason for standardization in this approach has not yet emerged? How can standardization of ethics measurement bridge the gap between principle and practice?

Purposive review firstly indicates that the principles-to-practice gap is not only due to insufficient regulation but how ethical constructs are defined, operationalized, measured and governed. Secondly, this reorients effort away from the production of additional isolated metrics, validation, comparison, and contextual calibration of measures. In sum, toward the comparatively neglected work of defining contested constructs and modelling trade-offs. Thirdly, it also identifies transparency and explainability measurement as a priority frontier, since the deficit there is more fundamental than the surplus in fairness.

For organizations, the framework cautions against treating any single metric as a sufficient demonstration of an ethical property and recommends that the choice of measure, and the handling of trade-offs, be documented and justified as a governable decision rather than a technical default. For policymakers and standards bodies, it clarifies that mandating quantifiable evidence is insufficient unless the validity and reliability of the underlying measures and the treatment of inter-principle trade-offs are also addressed; otherwise, regulation may incentivize measurement of whatever is convenient rather than of what matters.

Coordinated standardization, pursued as the three linked tasks set out above, is the institutional condition under which responsible AI could finally operate the commitments of ethical AI. Existing governance instruments illustrate the point: the NIST AI Risk Management Framework organizes risk work into

functions that include a distinct measure activity, and ISO/IEC 42001 institutionalizes AI management through process requirements (National Institute of Standards and Technology, 2023; International Organization for Standardization, 2023). Yet both remain largely procedural, presupposing rather than supplying the validated construct definitions and trade-off standards that the three linked tasks are meant to build.

CONCLUSION AND RECOMMENDATIONS

In conclusion, this paper argues that ethical AI and responsible AI carry a persistent gap which is fundamentally a measurement problem, instead of failure of regulation, organizational commitment. While ethical AI becomes the base of normative principles, responsible AI demonstrates, assesses and governs those principles in practice for AI systems. The literature review of this paper indicates that this measurement layer remains fragmented and unstandardized. While fairness is an overabundance of competing metrics, transparency and explainability limited evidence of construct validity, reliability, and standardized evaluation procedures. In addition, lack of standardized approaches for trade-offs among ethical principles limits businesses to implement responsible AI.

The contribution of this paper is to reframe the principles-to-practice gap as a measurement problem and to propose the three-task bridge developed in the proposed framework section, without which ethical commitments cannot be measured, compared, or governed.

This study carries valuable implications. Theoretically, it extends existing responsible AI literature by positioning measurement as the critical mechanism that connects normative ethics with organizational governance. Practically, the framework provides researchers, organizations, standards bodies, and policymakers with a structured foundation for designing more transparent evaluation processes, improving AI audits, and supporting the development of harmonized governance standards.

This study is subject to several limitations. First, it is conceptual in nature and does not empirically validate the proposed framework across real-world AI systems. Although an illustrative hypothetical operational scenario of an AI system to screen job applicants is mentioned, however in real word scenario validation at each stage actor, key activities and stage output are still to be validated through real case study. Second, the analysis focuses primarily on fairness, transparency, and explainability, while other principles such as privacy, robustness, sustainability, and human oversight warrant similar investigations. Finally, because ethical values vary across domains, cultures, and regulatory environments, the proposed framework should be interpreted as a flexible foundation rather than a universally applicable measurement standard.

Future research should therefore focus on empirically validating the proposed framework across different industries and AI applications, developing reliable measurement methods for underexplored principles such as transparency and explainability, and establishing standardized approaches for evaluating

trade-offs among competing ethical objectives. Collaborative efforts involving researchers, industry, standards organizations, and regulators will be essential to developing internationally accepted benchmarks and governance mechanisms.

Ultimately, responsible AI cannot be achieved through ethical AI principles alone. Ethical commitments become meaningful only when they can be measured, compared, challenged, and continuously improved. Advancing standardized and valid measurement practices is therefore a necessary step toward transforming ethical aspirations into accountable and trustworthy AI governance.

ACKNOWLEDGEMENTS

I would like to thank my mentor, Dr. Mike Smith, for his endless help, support, and guidance with my research *From Principle to Practice: Measurement as the Bridge Between Ethical AI and Responsible AI*, which brought it into its form.

REFERENCES

- Adcock, R., & Collier, D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, 95(3), 529–546. <https://doi.org/10.1017/S0003055401003100>
- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, 563(7729), 59–64. <https://doi.org/10.1038/s41586-018-0637-6>
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Benetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Blodgett, S. L., Lopez, G., Olteanu, A., Sim, R., & Wallach, H. (2021). Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1004–1015). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.81>
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 153–163. <https://doi.org/10.1089/big.2016.0047>
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv*. <https://doi.org/10.48550/arXiv.1702.08608>
- European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).

- Official Journal of the European Union, L 2024/1689.
<https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
- Floridi, L., & Cows, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.8cd550d1>
- Gyevnár, B., & Kasirzadeh, A. (2026). Bridging the gap in the responsible AI divides. Institute for Complex Social Dynamics, Carnegie Mellon University. <https://arxiv.org/pdf/2603.14495v1>
- Hagendorff, T. (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>
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems* 29 (pp. 3315–3323). Curran Associates.
<https://arxiv.org/abs/1610.02413>
- Infocomm Media Development Authority. (2026). Model AI governance framework for agentic AI. Government of Singapore.
- International Organization for Standardization. (2023). Information technology — Artificial intelligence — Management system (ISO/IEC Standard No. 42001:2023).
<https://www.iso.org/standard/81230.html>
- Jaakkola, E. (2020). Designing conceptual articles: Four approaches. *AMS Review*, 10(1–2), 18–26.
<https://doi.org/10.1007/s13162-020-00161-0>
- Jacobs, A. Z., & Wallach, H. (2021). Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 375–385). Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445901>
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- Johnson, D. G., & Verdicchio, M. (2025). The sociotechnical entanglement of AI and values. *AI & Society*, 40(1), 67–76. <https://doi.org/10.1007/s00146-023-01852-5>
- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)* (Article 43, pp. 43:1–43:23). Schloss Dagstuhl — Leibniz-Zentrum für Informatik.
<https://doi.org/10.4230/LIPIcs.ITCS.2017.43>
- Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10), 36–43.
<https://doi.org/10.1145/3233231>
- Maslej, N., Fattorini, L., Perrault, R., et al. (2025). The AI Index 2025 annual report. Stanford Institute for Human-Centered Artificial Intelligence. <https://hai.stanford.edu/ai-index/2025-ai-index-report>
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), Article 115.
<https://doi.org/10.1145/3457607>
- 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>

- Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11), 501–507. <https://doi.org/10.1038/s42256-019-0114-4>
- Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2024). Auditing large language models: A three-layered approach. *AI and Ethics*, 4(4), 1085–1115. <https://doi.org/10.1007/s43681-023-00289-2>
- Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., & Floridi, L. (2021). Ethics as a service: A pragmatic operationalisation of AI ethics. *Minds and Machines*, 31(2), 239–256. <https://doi.org/10.1007/s11023-021-09563-w>
- 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>
- Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., & Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys*, 55(13s), Article 295. <https://doi.org/10.1145/3583558>
- Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2021). Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), Article 7. <https://doi.org/10.1145/3449081>
- Reyes-Cruz, M., Gutiérrez-López, A., & García-Sánchez, F. (2025). Responsible AI: Conceptual boundaries, overlaps, and tensions with ethical AI. *AI and Ethics*, 6(1), 45–62. <https://doi.org/10.1609/aies.v8i3.36704>
- Rismani, S., Davis, L., Mingole, B., Rostamzadeh, N., Shelby, R., & Moon, A. (2025). Responsible AI measures dataset for ethics evaluation of AI systems. *Scientific Data*, 12, Article 1980. <https://doi.org/10.1038/s41597-025-06021-5>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Salaudeen, O., Reuel, A., Ahmed, A., Bedi, S., Robertson, Z., Sundar, S., Domingue, B., Wang, A., & Koyejo, S. (2025). Measurement to meaning: A validity-centered framework for AI evaluation. *arXiv*. <https://doi.org/10.48550/arXiv.2505.10573>
- 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>
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684), 677–680. <https://doi.org/10.1126/science.103.2684.677>
- Suárez-Álvarez, J., He, Q., Guenole, N., & D'Urso, D. (2026). Using artificial intelligence in test construction: A practical guide. *Psicothema*, 38(1), 1–12. <https://doi.org/10.70478/psicothema.2026.38.01>

- Timmermans, S., & Epstein, S. (2010). A world of standards but not a standard world: Toward a sociology of standards and standardization. *Annual Review of Sociology*, 36, 69–89. <https://doi.org/10.1146/annurev.soc.012809.102629>
- UAE Government. (2025). Dubai universal blueprint for artificial intelligence. <https://u.ae/en/about-the-uae/strategies-initiatives-and-awards/strategies-plans-and-visions/government-services-and-digital-transformation/dubai-universal-blueprint-for-ai>
- Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Article 440, pp. 1–14). Association for Computing Machinery. <https://doi.org/10.1145/3173574.3174014>
- Yang, J., Xie, W., & Ni, J. (2025). A framework for AI ethics literacy: Development, validation, and its role in fostering students' self-rated learning competence. *Scientific Reports*, 15, Article 38030. <https://doi.org/10.1038/s41598-025-21977-5>

ABOUT THE AUTHOR

Syed Mohammad Shayaan Hasnain is a Grade 12 student at The Westminster School, Dubai. His research interests lie in artificial intelligence governance and the measurement of ethical principles in computational systems, with a particular focus on how organizations translate normative commitments into accountable and auditable practice.