Making AI Risk Assessment More Objective: Addressing Sociotechnical Challenges

Matt A. Murphy
Independent Researcher, Denver, United States

Victoria Matthews
Independent Researcher, New York, United States

Abstract

Current AI risk assessment methods in industry rely heavily on subjective judgment. Yet recent U.S. AI policy frameworks, including Executive Order 14319 and America's AI Action Plan, mandate that AI systems be "free from ideological bias" and pursue "objective truth". While NIST provides systematic risk evaluation guidance, current AI governance frameworks are not designed to meet such objectivity requirements, instead offering flexibility that accommodates implementation across various contexts. Organizational AI risk practices rely heavily on subjective likelihood and impact scoring, compounded by subjectivity introduced when results are translated from technical teams to executives. Literature addressing this gap spans three disconnected streams: (1) AI governance frameworks relying on subjective risk assessment, (2) technical ML bias measurement focused on algorithmic fairness, and (3) organizational implementation approaches failing to translate governance principles into operational practice. Previous approaches lack links between risk analysis, objectivity requirements, and actionable guidelines, while policy frameworks remain too broad for practical application. Missing is research systematically bridging policy objectivity requirements with sociotechnical challenges of organizational risk assessment. To address this gap, we propose a framework transforming observable organizational factors into measurable risk indicators. Drawing from socio-technical systems theory and established risk taxonomies, we decompose traditional likelihood and impact metrics into specific, observable criteria replacing subjective estimates vulnerable to biases, addressing organizational tendencies to underestimate risks.

1 Introduction

The high-profile ethical failures of AI systems across industries underscore the urgency of robust and substantive governance and risk assessments. Amazon's AI-powered hiring system discriminated against women due to training data bias (Harvard Edmond J. Safra Center for Ethics, 2024). Zillow's home-pricing algorithm led to $304 million in losses and 25% workforce reduction when the system consistently overvalued properties (Florentine et al., 2022). Knight Capital suffered a $440 million loss in a single hour when a flawed trading algorithm made erroneous purchases across 150 stocks (Holistic AI, 2024). These failures highlight how subjective risk assessments that underestimate vulnerabilities can lead to catastrophic organizational and societal consequences. The occurrence of these failures despite organizations following standard risk assessment practices prompts many to ask the questions: when and why do AI risk assessments fail

Article presented at the Second IASEAI Conference (IASEAI'26).

to meaningfully capture risks? Based on recent U.S. AI policy frameworks, which mandate that AI systems be "free from ideological bias" and pursue "objective truth," the problem may lie in risk assessment subjectivity.

This paper introduces a novel 64-factor framework that attempts to systematically counter cognitive biases arising in AI risk assessment to meet emerging US policy objectivity standards. The framework bridges the critical gap between technical AI safety measures and practical organizational risk management through grounded, observable factor evaluation that reduces subjective judgment while maintaining stakeholder usability and regulatory compliance. Rather than simply identifying what to assess when it comes to AI risks (namely, their likelihood and impact), our approach provides systematic guidance on how to assess these dimensions through specific, verifiable organizational characteristics. This represents a fundamental shift from intuitive risk estimation toward evidence-based evaluation that can be documented, audited, and improved over time.

1.1 Recent US AI Policy Requirements Creating Objectivity Mandates

The landscape of AI governance has fundamentally shifted with recent U.S. policy developments that explicitly mandate objectivity in AI. Executive Order 14319, "Preventing Woke AI in the Federal Government," represents a watershed moment in federal AI policy. The order states unequivocally that "Americans will require reliable outputs from AI, but when ideological biases or social agendas are built into AI models, they can distort the quality and accuracy of the output" (The White House, 2025b).The complementary AI Action Plan builds upon this foundation by emphasizing the "[u]pdating" of "Federal procurement guidelines to ensure that the government only contracts with frontier large language model developers who ensure that their systems are objective and free from top-down ideological bias" (The White House, 2025a). The Plan explicitly states that AI systems must be "free from ideological bias and be designed to pursue objective truth rather than social engineering agendas."

These policy directives represent an unprecedented shift in federal AI governance. Where previous guidance from NIST and other agencies emphasized risk management as a flexible, context-dependent process (NIST, 2023), the new requirements demand concrete, demonstrable objectivity. While previous international policy frameworks on AI assessment, such as UNESCO's recommendations, provide a baseline for ethical principles, they likely remain too general for substantive organizational application beyond this baseline (UNESCO, 2021). This creates substantial implementation challenges for organizations, as the mechanisms for achieving and verifying "objectivity" in AI systems, remain poorly defined and up to interpretation (Binns, 2018).

1.2 The Cognitive Bias Crisis in Risk Assessment

The challenge of achieving objectivity in AI risk assessment is compounded by well-documented patterns of cognitive bias in organizational decision-making. Extensive research demonstrates that humans systematically underestimate risk probabilities and impacts due to cognitive biases and productivity incentives (Tversky & Kahneman, 1974; Sharot, 2011). These biases are not simply individual failures but emerge from predictable patterns in human cognition that affect even expert decision-makers (Gigerenzer, 2008). Optimism bias—overestimating positive outcomes and underestimating risks—is especially pervasive in organizations, where cognitive biases broadly distort decision-making and performance (Fasolo et al., 2025).

The detrimental influence of cognitive biases on decision-making and organizational performance is well established in management research (Fasolo et al., 2025). Recent systematic reviews of strategic decision-making confirm that "the strategic role of cognitive biases appears to be crucial when investigating the related impact on strategic decisions in times of environmental transformation" (Acciarini et al., 2021), making AI deployment (which represents precisely such a transformation) particularly susceptible to bias-driven risk assessment errors. Optimism bias leads teams to systematically underestimate the likelihood of failures, while anchoring effects cause initial risk estimates to unduly influence subsequent assessments (Shepperd et al., 2013). When

practitioners encounter an initial probability estimate—whether from a colleague, a preliminary analysis, or their own first impression—this anchor tends to constrain their subsequent judgments even when new information suggests different risk levels. Availability heuristics make recent or memorable incidents disproportionately influential in risk judgments, while confirmation bias leads assessors to seek information supporting pre-existing beliefs about system safety (Sunstein, 2002). These biases interact with organizational pressures to deploy systems quickly, creating systematic tendencies to minimize perceived risks to accelerate timelines (Emmons & Smalley, 2020).

The organizational context amplifies these cognitive biases. Teams developing AI systems often face strong incentives to characterize risks as manageable, as higher risk ratings may delay deployment or require additional investment in controls (Leveson, 2011). Project timelines, budget constraints, and competitive pressures create environments where acknowledging significant risks carries organizational costs. Research on organizational accidents demonstrates that the gradual acceptance of degraded safety standards, referred to as "normalized deviance", frequently precedes major failures (Vaughan, 1996). In AI contexts, this may manifest as teams gradually accepting higher levels of risk through repeated small decisions to proceed despite identified concerns.

1.3 Evidence for Better Risk Management Practices

Beyond policy mandates and documented cognitive biases, substantial evidence shows current AI risk management practices require improvement. Surveys of AI experts reveal deficiencies in risk management, with practitioners citing weaknesses in applying traditional frameworks to AI systems that exhibit novel failure modes (MIT Sloan Management Review, 2023). Research examining actual AI deployment failures reveals patterns of inadequate risk assessment preceding harmful outcomes (Raji et al., 2020; Florentine et al., 2022; Birkstedt et al., 2023). The Microsoft Tay chatbot incident revealed inadequate assessment of how malicious users might manipulate the learning algorithm (Palo Alto Networks, 2024), while multiple COVID-19 diagnostic AI tools failed in clinical practice because risk assessments did not identify fundamental flaws in training data (Florentine et al., 2022). These failures highlight the need for more structured approaches that guide teams toward comprehensive consideration of risks rather than relying on ad hoc judgments.

The rapid proliferation of AI applications across sectors has outpaced the development of specialized risk assessment expertise (Cath et al., 2018). Organizations deploying AI often lack personnel with deep experience in both AI technical capabilities and risk management methodologies, creating knowledge gaps that undermine assessment quality. A healthcare organization implementing AI for diagnostic support may have clinicians who understand medical risks but lack AI expertise, and data scientists who understand AI capabilities but lack risk management training. This expertise deficit makes it particularly important to develop frameworks that provide concrete guidance rather than relying primarily on expert judgment (Fjeld et al., 2020). Without structured approaches, organizations risk missing critical vulnerabilities that would be apparent to specialists but are not obvious to generalist teams conducting risk assessments.

2 Method

There are three core components to the risk assessment process we propose: a risk taxonomy, categories that group these risks, and the underlying material factors that help evaluate risk likelihood and severity. These components are all layers that make the risk assessment process more accessible for AI practitioners that undertake it.

2.1 Risk Taxonomy

The risk taxonomy found in Appendix A (full paper and appendices available at https://ssrn.com/abstract=6701598) is one of many risk taxonomies that have been developed for assessing AI risk. This particular taxonomy, which we developed for illustrative purposes, is geared towards organizations outside of the technology and AI industries, which may be using AI for their domain specific purpose. The risks captured are a subset of AI risks that arise in practical implementations of AI, and in practice would include more detailed descriptions to add precision.

They are primarily rooted in the ambit of longstanding risk frameworks, including Model Risk in financial services and broader Enterprise Risk Management, which is used across many contexts (Papadopoulos, 2015; Boultwood, 2021). The financial services sector has developed sophisticated approaches to model risk management over decades, driven by regulatory requirements and the high stakes of financial modeling errors (Federal Reserve SR 11-7, 2011). These frameworks recognize that models (whether statistical, machine learning, or rule-based) introduce risks through design limitations, implementation errors, and misuse in practice. Enterprise Risk Management (ERM) provides a broader organizational perspective on risk that encompasses strategic, operational, financial, and compliance dimensions (COSO, 2017). ERM frameworks emphasize the interconnected nature of risks and the importance of enterprise-wide coordination in risk identification and mitigation. Our taxonomy builds on these established approaches while adapting them to address AI-specific challenges such as opacity, emergent behaviors, and sociotechnical interactions.

Ultimately, this risk taxonomy serves to illustrate a sample spectrum of AI risks that may be relevant to business operations. It is our hope that it may be a starting place for how organizations can identify AI risks. The taxonomy is not intended to be exhaustive but rather to provide comprehensive coverage of the major risk domains that organizations may encounter when deploying AI systems (Brundage et al., 2020). Different organizations are likely to customize or extend the taxonomy based on their specific contexts, regulatory environments, and risk appetites, but the framework provides a foundation that captures core categories of AI-related risks across diverse deployment scenarios.

2.2 Risk Categories

These risks are then categorized by their shared characteristics or domain of relevance. Displayed in Appendix A (full paper and appendices available at https://ssrn.com/abstract=6701598), these risk categories are used to sort a list of observable factors related to AI into more usable subsets. The seven risk categories that reflect different domains where AI risks manifest are as follows:

- Security and Control Risks: Risks posed from the larger organizational frameworks and contexts that manage AI and allow it to function in safe, accountable, and secure ways
- Data Risks: Data integrity, ownership, poisoning, and evasion attacks that challenge a model's reliability
- Perception and Understanding Risks: Lack of data transparency and provenance to understand AI decision-making, as well as a lack of acute knowledge of AI's capabilities
- Ethical and Human Rights Risks: Negative or unintended consequences of AI development or deployment that affect individual rights, fairness, or societal well-being.
- Model Risks: Risks that arise from the performance of an AI model, based on the way it interacts with underlying components and data
- Business and Legal Risks: Legal and/or reputational harms that impact compliance with regulation or influence reputational standing
- Third Party Risks: Harms associated with the involvement of third parties (e.g., vendors) in creating, managing, or deploying AI or data solutions

These categories may differ by organization but are used to sort this paper’s primary factor taxonomy contribution for usability.

2.3 Observable Factors and Indicators

The observable factor framework is rooted in these risk categories and is the primary novel contribution of this paper. The 64 observable factors represent systematic decomposition of what

makes AI implementations risky in organizational practice, breaking down the commonly employed concepts of likelihood and impact to assess risk. Table 1 shows a shortened example of the factor framework, while the whole factor framework appears in Appendix B (full paper and appendices available at https://ssrn.com/abstract=6701598).

Table 1: Example Factor Taxonomy Entries

Risk Category	**Example Impact Factors**	**Example Likelihood Factors**
Security and Control Risks	System autonomy level; Dependency on continuous operation	Attack surface scope; Security incident frequency
Data Risks	Data sensitivity level; Data volume scope	Source data quality; Data handling complexity
Perception and Understanding	Decision transparency requirements; Stakeholder impact scope	Model interpretability level; Documentation depth
Ethical and Human Rights Risks	Demographic impact scope; Remediation difficulty	Population diversity; Historical bias incidents
Model Risks	Output criticality; Integration dependency	Model complexity level; Data drift frequency
Business and Legal Risks	Regulatory coverage scope; Compliance violation impact	Regulatory change frequency; Compliance framework maturity
Third Party Risks	Vendor dependency level; Contract scope/value	Integration complexity; Vendor stability issues

Each factor corresponds to specific characteristics that can be documented, measured, or verified through evidence. For instance, instead of asking practitioners to estimate "What is the likelihood of a bias-related incident?", we instead examine specific factors, such as whether bias testing has been conducted, if this is a novel application area, what demographic analysis results show for the diversity of populations affected by the model, and whether historical frequency of bias incidents have been measured and accounted for. These factors can all point us towards a clearer picture of likelihood for this particular risk. This approach transforms an ambiguous cognitive task into a structured evidence-gathering exercise (Simon, 1996). Rather than requiring practitioners to synthesize complex likelihood judgments from scratch, the framework breaks down the assessment into concrete questions about observable system and organizational characteristics. Research on judgment under uncertainty demonstrates that such decomposition improves accuracy by reducing the cognitive burden of holistic likelihood (or probability) estimation (Mellers et al., 2014). For example, when faced with the question "How likely is this AI system to exhibit bias?", practitioners must mentally integrate information about training data, model architecture, testing procedures, deployment context, and many other factors into their reasoning to arrive at a risk rating or score. This framework instead asks separate questions about each factor, reducing the risk that critical considerations will be overlooked or underweighted in a holistic judgment. Similarly, to separately

assess impact severity, we ask about the number of users and stakeholders that would be impacted by biased model determinations, as well as the inherent difficulty to remediate issues related to bias. In this sense, this factor framework decomposes likelihood and impact and goes a step further than where most current risk assessment approaches stop short.

The likelihood and severity of risk factors are obviously related, so how do we ensure that factors that fall in the likelihood category, as opposed to the impact severity category are not conflated? Though there is strong correlation between likelihood and impact, we attempt to address conceptual overlap in two ways. Firstly, we distinguish the fundamentally different questions that feed into determining likelihood, versus impact. Impact severity factors measure "if something goes wrong, how much damage occurs?", looking at scope, scale, and severity of consequences. Likelihood factors measure "what conditions make something more likely to go wrong?", examining vulnerabilities and likelihood indicators. A system can simultaneously have high potential impact (affecting millions of users) but low failure likelihood (due to robust controls), or vice versa. This independence allows the framework to guide different risk management strategies: impact informs what needs the strongest protection, while likelihood informs where to focus prevention efforts (Kaplan & Garrick, 1981).

Second, we break likelihood and impact severity into separate components that measure each dimension independently. This aims to prevent users of the framework from counting the same factor twice (as both a contributor to likelihood *and* a contributor to impact) while capturing its full risk profile. For instance, "system complexity" influences both how bad failures can be and how likely they are to occur. We split this into two distinct factors: "system autonomy level" (measures impact through how much damage occurs when autonomous systems fail without human oversight) and "model complexity level" (measures likelihood through how technical complexity increases chances of failure). This approach maintains analytical precision while avoiding double-counting, enabling better risk assessment and more targeted responses than treating complex concepts as single factors. It is of course not possible to completely eliminate factor overlap between impact and likelihood dimensions (Aven, 2016). However, when using additive scoring methods like ours, small overlaps have minimal impact on final scores and maintaining comprehensive risk coverage is more important than achieving perfect theoretical separation. This pragmatic approach acknowledges that real-world risk assessment requires balancing theoretical purity against practical utility. The goal is not mathematical perfection but rather systematic consideration of all relevant risk dimensions in a way that reduces cognitive burden and bias while maintaining sufficient analytical rigor to support sound decision-making.

2.4 Addressing Unstructured Assessment Processes

In a typical two-dimensional AI risk evaluation (often portrayed as a risk "matrix"), once risks are identified as part of a risk measurement exercise, they are given scores associated with both likelihood and impact (Acebes et al., 2024; Microsoft, 2024). This is where the process becomes more unstructured and relies on subjective human judgment of these two variables. Some broad structuring guidelines are typically given during a risk measurement exercise about what might affect likelihood or impact, but the evaluation process is mainly reliant on a combination of intuitive judgment and critical analysis of the AI use case context. Given the complex and emerging nature of AI governance, AI researchers or development teams may feel confused or ill-equipped to accurately rate these two variables. Research on AI practitioners' experiences with responsible AI frameworks reveals widespread uncertainty about how to operationalize abstract principles into concrete assessments (Rakova et al., 2021). Teams report struggling to translate high-level guidance about fairness, transparency, or safety into specific evaluation criteria applicable to their systems. A practitioner facing guidance to assess whether their system poses "high fairness risks" must determine what constitutes high risk, what evidence would support such a determination, and how to weigh different considerations that may point in different directions. This confusion is where this paper's primary methodological contribution may assist, by associating discrete, observable factors to underlying risk variables. The framework provides a structured pathway from abstract risk concepts to concrete organizational evidence, reducing the cognitive load on assessors while improving consistency across evaluations. Each question in the factor framework has a clear answer

based on available evidence, and the aggregate of answers provides a more systematic basis for the overall risk rating than an unstructured holistic judgment.

2.5 Materiality-Based Approach

Our methodology departs from traditional likelihood and impact estimation toward a materiality-based approach. Rather than asking teams to estimate the likelihood of AI failures, we systematically evaluate observable organizational characteristics and system factors indicating vulnerability to different risk types. This approach draws from risk assessment practices in financial model risk, nuclear safety, and critical infrastructure protection, where the focus shifts from predicting probabilities to identifying risk-contributing factors. The fundamental insight is that, while we cannot reliably predict when AI systems will fail, we can systematically identify conditions creating vulnerability. Materiality, in this context, represents the contextual significance of potential consequences—assessing not just harm magnitude (impact), but whether that harm matters enough to stakeholders, organizational objectives, regulatory bodies, or mission-critical functions to warrant action (Taleb, 2007).

In nuclear safety, the International Atomic Energy Agency's defense-in-depth strategy emphasizes identifying and mitigating vulnerabilities through multiple independent layers rather than relying solely on probabilistic calculations (IAEA, 2008). This approach recognizes that complex sociotechnical systems can fail in unanticipated ways that probabilistic risk assessments may not capture. The strategy evaluates materiality by considering which failures would be significant enough to threaten system integrity or public safety, structuring systems so no single failure can cause catastrophic outcomes—a principle highly applicable to AI systems whose failure modes may be difficult to enumerate in advance.

In financial services, there is a noted shift toward materiality-based evaluations of model risk to address business-line impacts in more tailored fashion (FDIC, 2017). Financial regulators increasingly emphasize understanding not just what would happen if models failed (impact magnitude), but whether those failures matter enough in the organizational and regulatory context to require disclosure, mitigation, or heightened controls, rather than demanding precise probability estimates for model failures.

This materiality lens may enable organizations to move beyond abstract impact ratings toward contextual risk prioritization. Our factor taxonomy is framed in these contextualized terms, prioritizing evaluation of which AI risks are significant enough to stakeholders, mission objectives, and regulatory obligations to warrant resource allocation, while acknowledging that conditions creating vulnerability (rather than precise failure probabilities) are what organizations can most reliably assess and manage.

2.6 Practical Implementation Guidance

The risk factors are ultimately used to underpin the likelihood and impact ratings in a risk score. A risk calculation may follow the structure: Risk Score = Likelihood × Impact × (1 - Control Effectiveness/5) where each component addresses different aspects of AI risk and are each rated on a 1-5 scale. We intend for our methodology to align with the practice of measuring likelihood and impact across respective Likert scales, often approached on a 1-5 or “Very Low” to “Very High” rating. The use of this kind of scale is not strictly necessary for the factors presented here to be useful, but it is a very common and useful theoretical foundation in risk practices (Pescaroli et al., 2020). Ideally, 1-5 ratings for likelihood, impact, and control effectiveness are assessed primarily using subject matter expertise and the context of the individual AI use case. However, to augment these unstructured parts of the assessment, a rule of thumb can be employed for converting the presence of factors from our list to the 1-5 rating.

Each factor (each bullet in the factor framework) that is considered a relevant contributor to a risk for an AI use case may yield one increased level of risk in the 5-point scale for a risk in that category (starting from 1 - Very Low Risk). For example, an AI use case must consider "Data Risk - Sensitive Data Exposure". There are at least three relevant impact factors in the framework for this risk (high

number of users, very sensitive data level, very valuable data), which means that the impact severity rating is at least a "4 or High" (1 + 3 relevant factors). This is of course not rigid but is a helpful guideline to frame the factors so that they are useful for the risk assessment. The approach provides a structured starting point while allowing assessors to apply contextual judgment when factors are partially present or when interactions between factors warrant adjustment (Aven & Renn, 2009). The guideline allows organizations to weight partially applicable factors, ensuring that nuanced strengths or weaknesses inform risk scoring without rigid formulas.

The categorization of the various factors serves practical implementation purposes by allowing different organizational functions to focus on risk categories most relevant to their expertise while ensuring systematic coverage across risk domains. This division of labor approach aligns with research on boundary objects that facilitate coordination across different professional communities (Star & Griesemer, 1989). Security teams can assess security and control risk factors using their specialized knowledge of threat landscapes and defensive measures. Data scientists can evaluate model and data risk factors based on their understanding of algorithmic behavior and data quality issues. Legal and compliance personnel can assess business and legal risk factors informed by their knowledge of regulatory requirements and organizational obligations. This allows the risk factor taxonomy to be distributed if organizational resources allow, remaining inclusive of specialized expertise, when possible, while still equipping non-experts with explicit indicators to measure.

3 Analysis: How This Evolves AI Risk Evaluation

Our factor-based methodology changes AI risk evaluation by providing more structured guidance for what has traditionally relied on unstructured human judgment during the process of risk assessment. Traditional risk assessment methods rely heavily on the reliability of data used and the skills and expertise of individuals conducting assessments, often leaving teams uncertain about what factors to consider or how to evaluate them consistently (Aven, 2012). This represents a classic sociotechnical challenge: how to design risk assessment frameworks that effectively support human decision-making within organizational contexts while accounting for cognitive limitations and structural constraints (Baxter & Sommerville, 2011; Kudina & van de Poel, 2024). Though the complete elimination of residual subjectivity in AI risk evaluation is not possible, our approach moves toward greater objectivity by offering concrete, observable criteria that reduce the cognitive burden (and variance) on individual assessors.

The shift from holistic judgments to structured factor evaluation addresses several documented problems in risk assessment. First, it reduces the influence of cognitive biases by breaking down complex judgments into simpler component evaluations less susceptible to systematic errors. Second, it improves consistency across assessors and over time by providing explicit evaluation criteria rather than relying on tacit knowledge or individual interpretation. Third, it creates accountability through documentation of the evidence supporting each factor assessment. Fourth, it facilitates learning by making the reasoning behind risk ratings explicit and reviewable. Organizations can examine past assessments, compare them to actual outcomes, and refine their evaluation processes based on this feedback in ways that are difficult when assessments rely primarily on undocumented intuitive judgments. This shift helps transform what may be an overwhelming cognitive task into a more manageable systematic process. The structured approach aligns with principles of cognitive task analysis and decision support systems design, which emphasize breaking complex tasks into manageable components and providing scaffolding for systematic evaluation (Klein, 2008).

2.7 Improving Information Flow

Our methodology improves information flow between technical teams and business decision-makers. The challenge is that technical teams understand AI capabilities but struggle to communicate meaningfully to executives, while executives must make deployment decisions, often without the full technical expertise needed to evaluate risk claims. The factor-based approach allows technical teams to present evidence in terms non-technical decision-makers can understand and verify. Executives can examine documented evidence for each factor—deployment scale, user

demographics, testing results, known limitations—and make informed decisions about risk acceptance based on organizational risk appetite. This transparency supports better decision-making while maintaining appropriate accountability, particularly critical for AI systems where deployment decisions require executive approval despite potential lack of technical AI expertise (Zerilli et al., 2019).

2.8 Reducing Cognitive Bias Impact

The structured factor-based approach directly addresses cognitive biases undermining traditional risk assessment. By requiring systematic examination of specific, predefined factors, the framework reduces availability bias—the tendency to overweight easily recalled examples (Tversky & Kahneman, 1973). Explicitly breaking likelihood and impact into measurable criteria minimizes anchoring effects by guiding assessors to evaluate probabilities systematically rather than relying on initial reference points (Mellers et al., 2014). Initial "low risk" impressions cannot persist unchallenged; the assessment requires examining specific evidence for each factor. The framework protects against optimism bias by requiring documentation of risk-increasing factors even when overall impressions are positive.

2.9 Creating Organizational Accountability

The factor-based approach creates clearer accountability by making risk rating bases explicit and auditable. When risk scores rely on subjective probability estimates, retrospective review is difficult. The observable factor approach allows examining whether documented evidence genuinely supports claimed factor presence or absence (Power, 2007). This serves multiple functions: enabling independent validation, supporting governance compliance, and facilitating organizational learning through post-incident reviews examining whether risk factors contributing to failures were appropriately identified. Explicit documentation creates incentives for careful assessment, as assessors face reputational consequences for careless evaluations (Hood et al., 2001). This accountability mechanism counteracts organizational pressures toward optimistic assessments.

2.10 Addressing AI-Specific Challenges

The framework addresses challenges specific to AI risk assessment that traditional frameworks struggle with. Rapid AI development means likelihood estimates based on historical data quickly become outdated (Brundage et al., 2020). The factor-based approach accommodates this dynamism by focusing on observable current characteristics rather than potentially obsolete statistical trends. New AI capabilities lacking historical data can still be evaluated through observable factors: user impact, data sensitivities, testing conducted, controls in place.

AI systems' complexity and opacity make predicting failure modes difficult. The framework acknowledges this uncertainty by assessing vulnerability indicators rather than demanding precise probability estimates. This aligns with risk analysis scholarship emphasizing that for complex systems, understanding what makes failures possible is more valuable than calculating how likely particular scenarios are (Aven, 2010).

2.11 Supporting Risk Communication and Decision-Making

The structured approach improves risk communication by providing concrete, verifiable claims rather than abstract likelihood estimates. When technical teams report "high risk," executives can examine specific evidence-based statements: user numbers affected, data sensitivity, model architecture novelty, monitoring capabilities, regulatory requirements. These concrete factors provide meaningful decision-making information. The framework also supports risk comparison across multiple AI systems, enabling meaningful comparison by examining which systems share risk factors and where they differ.

3 Conclusion

This paper presents a novel approach to AI risk assessment addressing critical gaps in current practice. By decomposing traditional risk dimensions into 64 observable factors, the framework provides concrete guidance reducing reliance on subjective likelihood and impact judgments. The framework challenges the notion that risk assessment should conceptually stop at the intuitive measurement of likelihood and impact, instead taking that assessment a step further by systematically decomposing these dimensions into observable, verifiable factors. This responds to policy mandates for objectivity while addressing well-documented cognitive biases undermining assessment quality. The framework represents one of the first systematic attempts to operationalize recent U.S. AI policy requirements into practical organizational guidance. Grounding evaluation in observable characteristics, the framework advances objectivity not by eliminating subjective judgment—which is neither possible nor desirable—but by reducing unnecessary subjectivity through systematic, evidence-based assessment. The methodology offers key advantages: transforming what might be overwhelming cognitive tasks into structured evidence-gathering, creating accountability through auditable documentation, facilitating communication across stakeholder groups through shared terminology, and accommodating AI-specific challenges traditional frameworks struggle to address.

However, important limitations remain. The framework requires empirical validation through deployment in diverse organizational contexts to assess impact on assessment quality and outcomes. While building on established principles and addressing documented weaknesses, its effectiveness in improving actual outcomes, such as reducing incidents, supporting better decisions, enabling effective mitigation, requires systematic evaluation. The framework will require ongoing refinement as AI capabilities evolve. Interaction effects between factors warrant further investigation, as some combinations may create synergistic risks requiring non-linear approaches. Future research should examine how organizations adapt the framework to illuminate best refinement opportunities.

The broader challenge of striving for greater objectivity and accuracy in AI evaluations extends beyond methodology to encompass organizational culture, incentive structures, and governance processes. No framework substitutes for organizational commitment to responsible use or deployment of technologies. However, by providing concrete tools making systematic assessment more feasible, this work contributes a practical infrastructure to support responsible AI governance. As AI systems become more capable and widely deployed, the stakes of assessment failures continue rising. Improving evaluation quality and consistency is prerequisite for maintaining public trust and ensuring AI systems serve human welfare. This framework offers one component of a broader solution, demonstrating that more systematic approaches are both possible and practical.

Ultimately, framework success depends on adoption and implementation by organizations developing and deploying AI systems. The framework prioritizes practical feasibility alongside theoretical rigor, recognizing that overly complex frameworks will not improve practice regardless of conceptual elegance. By building on familiar structures while adding systematic guidance, the approach seeks to lower adoption barriers while delivering meaningful assessment quality improvements. Whether this balance succeeds requires real-world testing, but the urgent need for better AI risk assessment practices justifies experimentation with systematic alternatives to current subjective approaches.

References

Acciarini, C., Brunetta, F., & Boccardelli, P. (2021). Cognitive biases and decision-making strategies in times of change: A systematic literature review. *Management Decision, 59*(3), 638–652.

Acebes, F., González-Varona, J. M., López-Paredes, A., & Pajares, J. (2024). Beyond probability-impact matrices in project risk: A quantitative methodology for risk prioritisation. *Humanities and Social Sciences Communications, 11*, 670. https://doi.org/10.1057/s41599-024-03180-5

Aven, T. (2010). On how to define, understand and describe risk. *Reliability Engineering & System Safety, 95*(6), 623–631.

Aven, T. (2012). The risk concept—historical and recent development trends. *Reliability Engineering & System Safety, 99*, 33–44.

Aven, T. (2016). Risk assessment and risk management: Review of recent advances on their foundation. *European Journal of Operational Research, 253*(1), 1–13.

Aven, T., & Renn, O. (2009). On risk defined as an event where the outcome is uncertain. *Journal of Risk Research, 12*(1), 1–11.

Baxter, G., & Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering. *Interacting with Computers, 23*(1), 4–17.

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of Machine Learning Research, 81*, 149–159.

Birkstedt, T., Minkkinen, M., Tandon, A., & Mäntymäki, M. (2023). AI governance: Themes, knowledge gaps and future agendas. *Internet Research, 33*(7), 133–167. https://doi.org/10.1108/INTR-01-2022-0042

Boultwood, B. (2021, July 16). *How to develop an enterprise risk taxonomy.* GARP. https://www.garp.org/risk-intelligence/culture-governance/how-to-develop-an-enterprise-risk-taxonomy

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., … & Anderljung, M. (2020). Toward trustworthy AI development: Mechanisms for supporting verifiable claims. *arXiv preprint* arXiv:2004.07213.

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.

Committee of Sponsoring Organizations of the Treadway Commission (COSO). (2017). *Enterprise risk management: Integrating with strategy and performance.*

Emmons, D. L., & Smalley, D. E. (2020). Mitigating cognitive biases in risk identification: Practitioner checklist for the aerospace sector. *PMC.* https://pmc.ncbi.nlm.nih.gov/articles/PMC7398041/

Fasolo, B., Misuraca, R., & McClelland, G. H. (2025). Individual differences in adaptive decision making: Implications for context effects and debiasing. *Organizational Behavior and Human Decision Processes.*

Federal Deposit Insurance Corporation (FDIC). (2017). *Guidance on model risk management* (Financial Institution Letter FIL-22-2017).

Federal Reserve. (2011). *Supervisory guidance on model risk management (SR 11-7).* Board of Governors of the Federal Reserve System.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication, 2020-1.*

Florentine, S., Perkins, B., & White, S. K. (2022). 11 famous AI disasters. *CIO.* https://www.cio.com/article/190888/5-famous-analytics-and-ai-disasters.html

Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science, 3*(1), 20–29.

Harvard Edmond J. Safra Center for Ethics. (2024). *Into the abyss: Examining AI failures and lessons learned.* https://www.ethics.harvard.edu/blog/post-8-abyss-examining-ai-failures-and-lessons-learned

Holistic AI. (2024). *An exploration of AI harms: The need for AI risk management.* https://www.holisticai.com/blog/exploration-of-ai-harms

International Atomic Energy Agency (IAEA). (2008). *Defence in depth in nuclear safety (INSAG-10).*

Kaplan, S., & Garrick, B. J. (1981). On the quantitative definition of risk. *Risk Analysis, 1*(1), 11–27.

Klein, G. (2008). Naturalistic decision making. *Human Factors, 50*(3), 456–460.

Kudina, O., & van de Poel, I. (2024). A sociotechnical system perspective on AI. *Minds and Machines, 34*(3), Article 21. https://doi.org/10.1007/s11023-024-09680-2

Leveson, N. (2011). *Engineering a safer world: Systems thinking applied to safety.* MIT Press.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., & Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science, 25*(5), 1106–1115. https://doi.org/10.1177/0956797614524255

Microsoft. (2024, March 29). *How to create a risk assessment matrix.* Microsoft 365 Life Hacks. https://www.microsoft.com/en-us/microsoft-365-life-hacks/organization/risk-assessment-matrix

MIT Sloan Management Review. (2023). *AI-related risks test the limits of organizational risk management.* https://sloanreview.mit.edu/article/ai-related-risks-test-the-limits-of-organizational-risk-management/

National Institute of Standards and Technology. (2023, January 26). *AI risk management framework.* https://www.nist.gov/itl/ai-risk-management-framework

Palo Alto Networks. (2024). *AI risk management framework.* https://www.paloaltonetworks.com/cyberpedia/ai-risk-management-framework

Papadopoulos, P. (2015, June 24). *Open risk taxonomy (OpenRisk white paper).* OpenRisk. https://www.openriskmanagement.com/wp-content/uploads/2017/02/OpenRiskWP04_061415.pdf

Pescaroli, G., Velázquez, O., Alcántara-Ayala, I., Galasso, C., Kostkova, P., & Alexander, D. (2020). A Likert scale-based model for benchmarking operational capacity, organizational resilience, and disaster risk reduction. *International Journal of Disaster Risk Science, 11*(3), 404–409. https://doi.org/10.1007/s13753-020-00276-9

Power, M. (2007). *Organized uncertainty: Designing a world of risk management.* Oxford University Press.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., … & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency,* 33–44.

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), 1–23.

Sharot, T. (2011). The optimism bias. *Current Biology, 21*(23), R941–R945. https://doi.org/10.1016/j.cub.2011.10.030

Shepperd, J. A., Klein, W. M., Waters, E. A., & Weinstein, N. D. (2013). Taking stock of unrealistic optimism. *Perspectives on Psychological Science, 8*(4), 395–411.

Simon, H. A. (1996). *The sciences of the artificial* (3rd ed.). MIT Press.

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, translations, and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. *Social Studies of Science, 19*(3), 387–420.

Sunstein, C. R. (2002). *Risk and reason: Safety, law, and the environment.* Cambridge University Press.

Taleb, N. N. (2007). *The Black Swan: The impact of the highly improbable.* Random House.

The White House. (2025a, July). *America's AI action plan.* https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

The White House. (2025b, July 23). *Preventing woke AI in the federal government.* https://www.whitehouse.gov/presidential-actions/2025/07/preventing-woke-ai-in-the-federal-government/

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131.

UNESCO. (2021, November 23). *Recommendation on the ethics of artificial intelligence.* https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

Vaughan, D. (1996). *The Challenger launch decision: Risky technology, culture, and deviance at NASA.* University of Chicago Press.

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology, 32*(4), 661–683.