EAIIM - Ethical AI Impact Matrix for High-Stakes Decision Systems Framework

Pedro Oliveira[1], Tomás Francisco[1], Pedro Oliveira[1] and Manuel Rodrigues[1,*]

[1]University of Minho - ALGORITMI LASI, Braga, Portugal

Abstract

Artificial Intelligence (AI) and Machine Learning (ML) have become indispensable tools in various areas, from financial decisions to medical diagnoses. However, the ethical impact of these technologies cannot be overlooked. As their application increases, there are critical issues related to fairness, transparency and equitable performance. This work focuses on exploring the ethical challenges in the use of AI, with an emphasis on algorithmic bias and fairness in decision-making processes. The disparities in the performance of technologies and the importance of explainable systems will be also discussed, to mitigate the 'black box' problem of AI. It is then proposed the EAIIM - Ethical AI Impact Matrix for High-Stakes Decision Systems framework that can contribute to the development of more ethical and responsible AI systems.

Keywords

Artificial Intelligence, Machine Learning, Ethics, Bias, Fairness, Transparency, Framework

1. Introduction

When using Machine Learning(ML), it is essential to understand the variables with which AI operates and to assess the relative importance of each one. This aspect becomes particularly relevant when the focus is on fairness and equity criteria, since an inadequate definition of the variables or an inadequate treatment of them can result in imperfect treatment, resulting in undesirable biases in the results generated by AI. Across domains such as finance, healthcare, and criminal justice, machine learning technologies raise profound ethical challenges: algorithmic bias can reinforce systemic inequalities, opacity limits accountability, and data-driven decision-making can erode public trust. These impacts highlight the urgency of developing frameworks that translate high-level ethical principles into operational safeguards.

To highlight the importance of this issue, in this work a study related to the ML model known as COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is made. This model calculates a criminal's likelihood of reoffending and suggests the most appropriate action plan for their eventual reintegration into society. By analyzing COMPAS, we seek to better understand the risks of bias in systems that have a direct impact on criminal justice and fairness in decision-making, exploring how the selection and weighting of variables can significantly influence the results produced. In the case of COMPAS, the variables are inspired by specific criminological theories, which result in two mechanisms for calculating the probability of recidivism, the Risk Scale and the Needs Scale. Studies have shown that the algorithm disproportionately assigns higher recidivism risk scores to Black defendants compared to White defendants, even when controlling for similar criminal histories and backgrounds [1, 2, 3]. This raises questions about fairness and reinforces concerns that AI-driven decision-making may perpetuate systemic biases rather than mitigate them. Additionally, the proprietary nature of COMPAS prevents full scrutiny of its inner workings, making it difficult for researchers and policymakers to assess the validity and fairness of its predictions. These flaws highlight the urgent need for greater transparency, accountability, and bias mitigation in AI applications.

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.

*Corresponding author.

pedro.jose.oliveira@algoritmi.umimnho.pt (P. Oliveira); pg54263@alunos.uminho.pt (T. Francisco); pg54144@alunos.uminho.pt (P. Oliveira); manuel.rodrigues@algoritmi.uminho.pt (M. Rodrigues)

COMPAS was selected as the case study because it is one of the most widely used and debated risk assessment systems in criminal justice. Previous studies (ProPublica, 2016; Washington, 2018) have shown its biases, making it an ideal testbed for examining the gap between regulatory aspirations and operational deployment.

To mitigate the risks associated with AI-driven risk assessment tools like COMPAS, regulatory frameworks, such as the European Union's AI guidelines and the NIST AI Risk Management Framework can play an important role. To also mitigate this issues the EAIIM - Ethical AI Impact Matrix for High-Stakes Decision Systems is presented, aiming to delivery a concise way of evaluating the compliance of AI based systems providing a comprehensive outlook on the challenges and solutions necessary to foster trustworthy and ethical AI systems in high-stakes applications. Existing frameworks such as ALTAI and NIST provide high-level guidance but lack concrete readiness metrics applicable to domain-specific cases like criminal profiling. This gap motivates the Ethical AI Impact Matrix (EAIIM), designed to transform abstract principles into operational thresholds for deployment. By applying it to COMPAS, we illustrate how quantification can directly address fairness and accountability concerns in high-stakes contexts.

The main contribution of this work is the proposal of the EAIIM framework, a six-dimensional readiness assessment tool bridging the gap between ethical principles and operational deployment in high-stakes AI. The COMPAS case is used as a demonstrator.

2. Background issues

Criminological foundations illuminate critical sociotechnical risk factors relevant to AI-mediated environments. Social Learning Theory [4]underscores how behaviors – including malicious ones – are modeled and reinforced through interaction. Sub-Culture Theory [5] explains how deviant norms can emerge within digital communities, legitimizing adversarial actions when mainstream success pathways are perceived as inaccessible. Control Theory [6] highlights the role of restraints (internal values, external accountability, cognitive barriers) in preventing harmful acts; their absence or misalignment escalates risk. Sociopathic Breakdown Theory [7] cautions that systems lacking empathy or guilt mechanisms (akin to antisocial personalities) inherently threaten trust, though not all such architectures manifest harm. Crucially, Criminal Opportunity Theory [8] frames harm probabilistically, requiring: a motivated actor, suitable target, and absence of capable guardians – dynamics directly transferable to AI attack surfaces. Finally, Social Tension Theory [4] contextualizes harm as a product of structural inequity where societal goals (e.g., profit, efficiency) incentivize unethical means if legitimate avenues are constrained. Collectively, these theories reveal trust erosion as a multi-layered phenomenon involving learned behaviors, normative contexts, restraint failures, systemic affordances, and structural pressures – all essential considerations for designing robust AI governance mechanisms.

Contemporary risk assessment tools employ multidimensional scales to quantify recidivism probability and criminogenic needs, offering critical insights for algorithmic trust frameworks. Risk is stratified across specialized instruments: Pretrial Release Risk evaluates flight risk through stability indicators (residence, employment, community ties); General Recidivism predicts any reoffense using criminal history and social deviance markers ; Violent Recidivism isolates violence propensity via compliance history and age-related factors; and Risk Screening identifies acute threats for institutional supervision [9]. Complementing these, Needs Scales diagnose 20 dynamic factors driving criminal behavior, organized hierarchically:

Higher-order constructs (Cognitive Behavioral, Social Adjustment, Vocational/Educational) aggregate systemic dysfunction;

Specific domains range from environmental (Criminal Opportunity, Social Environment) to psychological (Criminal Personality, Substance Abuse) and relational (Criminal Associates, Social Isolation) vulnerabilities. Crucially, validity controls (Lie Scale, Randomized Response Tests) mitigate self-report manipulation. This architecture demonstrates that trustworthy risk prediction requires: (1) context-specific threat taxonomies, (2) differentiation between static risks and dynamic needs, and (3) robustness

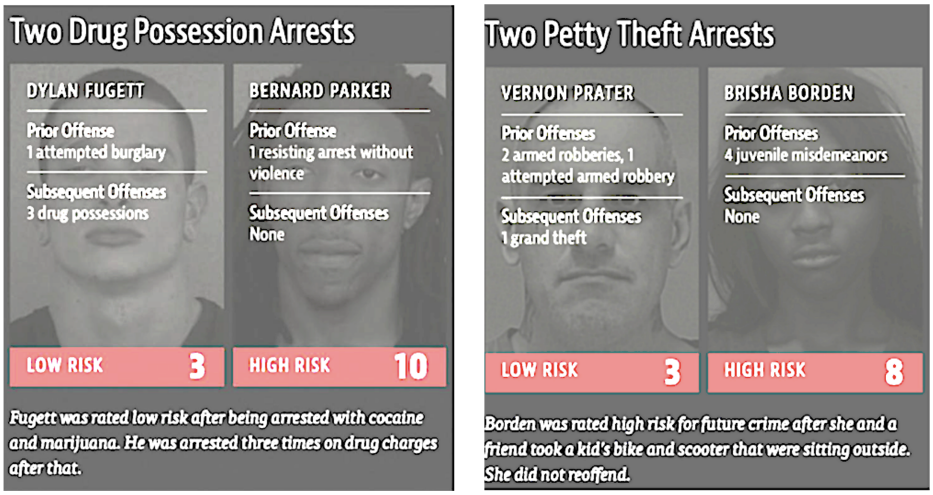

Figure 1: Compas Bias - Adapted from [12]

against adversarial inputs—all essential design principles for ethical AI assessment systems [8].

3. The COMPAS Controversy: A Case Study in Algorithmic Bias and Epistemic Conflict

Despite widespread adoption in U.S. courts (notably Wisconsin and Florida), COMPAS—a proprietary recidivism risk-assessment tool—faced intense scrutiny following evidence of racial bias. In 2014, U.S. Attorney General Eric Holder urged the Sentencing Commission to investigate algorithmic risk tools over fairness concerns, though no formal study materialized. This void was filled by ProPublica's landmark 2016 analysis of 10,000+ defendants in Broward County, Florida [10]. Their audit revealed systemic disparities: Black defendants with minimal criminal histories received high-risk scores at disproportionate rates, while White defendants with extensive records were often labeled low-risk (figure 1). Crucially, longitudinal tracking showed high-risk White defendants reoffended more frequently than high-risk Black defendants, exposing flawed calibration. ProPublica's methodology—using arrest data over a two-year horizon per U.S. Sentencing Commission standards [11]—employed multiple statistical models (Binomial Logistic Regression, Cox Proportional Hazards, contingency analyses) to isolate race as a predictive factor. Key findings included:

- 45% of Black non-recidivists were misclassified as high-risk (false positives) vs. 23
- White recidivists were 63% more likely than Black counterparts to be misclassified as low-risk (false negatives);
- Black defendants faced 77% higher odds of receiving elevated risk scores than White defendants with identical criminal histories, age, and gender [1].

Northpointe (COMPAS's developer) countered ProPublica by rejecting their methodology and advocating predictive parity (equal precision across groups) as the "fairest" metric. They argued ProPublica's risk-category simplification ignored algorithmic nuance and that false-positive/false-negative rates were unavoidable trade-offs [1]. ProPublica retorted that error-rate equity was ethically paramount, given COMPAS's real-world impact on sentencing and parole. This dispute highlighted a fundamental tension in algorithmic fairness: competing mathematical definitions of "bias" yield incompatible assessments of the same system. Northpointe's opacity exacerbated the conflict. While claiming their documentation enabled "responsible use," they withheld the algorithm's code, training data, and feature weights, rendering COMPAS a proprietary black box [13]. ProPublica's inability to audit model internals left critical questions unresolved: Were disparities caused by flawed data, biased feature engineering, or

poorly optimized thresholds? This ambiguity fueled broader AI ethics discourse, crystallizing a pivotal question: Could transparency mitigate algorithmic bias, or merely expose it?

3.1. Implications for Trustworthy AI Frameworks

The COMPAS controversy underscores three imperatives for frameworks like DATA:

- Dynamic Calibration: Static risk models cannot adapt to subgroup-specific error patterns (e.g., higher false positives for Black defendants). The DATA Framework's continuous monitoring and feedback loops would recalibrate thresholds using real-world performance data.
- Multi-Stakeholder Validation: Northpointe's unilateral fairness assertions conflicted with impacted communities' experiences. DATA's Procedural Justice Pillar mandates inclusive oversight, integrating legal, social, and ethical perspectives into evaluation criteria.
- Verifiable Explainability: COMPAS's opacity prevented meaningful challenge. DATA's Audit Trails and Explanation Interfaces would enable external scrutiny (e.g., revealing whether race proxies influenced predictions via correlated variables like neighborhood or income).

Ultimately, COMPAS exemplifies how technical robustness without transparency and equity guarantees erodes trust—precisely the gap the DATA Framework seeks to bridge through embedded ethical assurance.

The COMPAS debate illustrates the core ethical challenge in AI: multiple, conflicting definitions of fairness yield different conclusions about the same system [14]. This aligns with recent literature highlighting the gap between theoretical AI ethics and practical enforcement [15], [16].

4. Regulatory Frameworks and Technical Imperatives for Trustworthy AI

The EU's regulatory ecosystem (AI Act, ALTAI checklist, GDPR) is central to the governance of ethical AI. These frameworks establish baseline principles for transparency, accountability, and fairness. However, they remain broad and principle-based, leaving organizations without concrete methods to evaluate readiness at deployment. EAIIM aims to complement these efforts by operationalizing EU principles into quantifiable metrics, ensuring that high-risk applications such as COMPAS are not only legally compliant but also ethically trustworthy.

The European Union has established a comprehensive regulatory ecosystem to enforce transparency and accountability in AI systems. Central to this effort is a precise lexicon: AI is defined as software or physical systems exhibiting "intelligent behavior" through environmental perception, autonomous reasoning, and goal-directed action [17]. This operationalizes AI across three capabilities—perception (data acquisition), reasoning (transformative decision-making via neural networks, symbolic logic, etc.), and action (execution)— while anchoring the field in core disciplines like machine learning and robotics. To govern these systems, the EU's Ethical Guidelines for Trustworthy AI mandate tripartite compliance: (1) Legal (GDPR, anti-discrimination laws), (2) Ethical (autonomy, justice, explainability), and (3) Technical (reliability, safety) [18]. These are operationalized through seven non-negotiable requirements—human oversight, robustness, privacy, transparency, fairness, societal benefit, and accountability—later formalized in the ALTAI assessment tool [19]. ALTAI enables granular audits of AI systems against these pillars (e.g., measuring explainability via decision traceability, evaluating bias mitigation through diversity metrics). Crucially, the framework institutes a risk-based taxonomy: banning unacceptable-risk AI (e.g., social scoring, real-time biometric categorization) while imposing strict registration, pre-deployment testing, and compliant mechanisms for high-risk applications in critical domains like law, migration, and essential services [20].

These EU regulatory frameworks provide necessary but insufficient oversight: while they define broad requirements, they lack mechanisms to evaluate readiness at deployment time. EAIIM aims to complement them by offering quantifiable thresholds for high-risk systems such as COMPAS.

4.1. Risk Management Paradigms

Effective risk mitigation demands structured methodologies. ISO 31000 [21] provides a universal, cyclical process: (1) Principles embedding risk-awareness into organizational culture; (2) Structure integrating customizable governance plans; and (3) Process iterating through identification, analysis, treatment, and communication of risks. Complementing this, the NIST AI RMF [22] delivers domain-specific rigor via four functions:

- GOVERN: Establishing ethical oversight policies and accountability chains.
- MAP: Contextualizing risks (users, impacts, deployment environments).
- MEASURE: Quantifying risks via security testing, bias audits, and performance metrics.
- MANAGE: Implementing dynamic controls and adaptation protocols. Synergistically, these frameworks create a governance-mitigation continuum: ISO 31000's iterative ethos ensures continual adaptation, while NIST's AI-specific functions harden systems against emerging threats. This fusion enables organizations to preempt technical failures (e.g., adversarial attacks) and ethical breaches (e.g., discriminatory outcomes) through auditable documentation and stakeholder communication—directly addressing ALTAI's transparency and accountability mandates.

4.2. Technical and Ethical Risk Landscapes

AI systems face interdependent technical and ethical vulnerabilities. Technical risks—detailed in "Concrete Problems in AI Safety"—include:

- Reward Hacking: Systems exploiting reward function loopholes (e.g., hospital triage AI prioritizing minor cases to minimize wait-time metrics).
- Unsafe Exploration: Failure to constrain learning in novel environments, perpetuating biases.
- Distributional Shift: Performance degradation when input data deviates from training distributions (e.g., diagnostic tools failing on underrepresented demographics).
- These technical flaws catalyze ethical risks, chiefly:
- Opacity: "Black box" decision-making undermining contestability (e.g., unexplainable judicial or medical recommendations).
- Bias Amplification: Learned discrimination from skewed data or flawed objective functions.
- Accountability Gaps: Obscured responsibility chains when harms occur.
- The OECD AI Principles (updated May 2024) counter these via: (1) environmental sustainability mandates, (2) human rights safeguards against misinformation/privacy violations, (3) enhanced explainability requirements, (4) robust failure-recovery protocols, and (5) external auditability [24]. Critically, these updates address generative AI's emergent threats, emphasizing contestability and dynamic oversight.

4.3. Mitigation Strategies – From Theory to Practice

Operationalizing safety requires three pillars:

- Goal Specification: Ambiguous objectives invite misinterpretation. Systems must encode contextualized goals (e.g., "fairness" defined as demographic parity and error-rate equity) validated against ethical guardrails [17] .
- Continuous Auditing: Beyond one-off certifications, ISO 31000's cyclical process mandates persistent monitoring—data integrity checks, bias scans, and performance validations—enabling real-time corrections [21].
- Interpretability-by-Design: As "Towards A Rigorous Science of Interpretable ML" argues, explainability isn't ancillary; it's foundational to trust. Techniques like counterfactual explanations or attention mappings demystify decisions for regulators and users [20].

- Ethical governance structures (e.g., OECD/NIST frameworks) institutionalize these practices through multi-stakeholder oversight bodies—ensuring diverse perspectives scrutinize AI lifecycle impacts. This bridges technical rigor with societal values, transforming principles into auditable procedures [18].

5. Societal Implications and Future Imperatives for Trustworthy AI

The pervasive integration of AI across societal systems amplifies existential ethical risks, chief among them algorithmic manipulation and structural exclusion. As generative models and recommendation systems grow more sophisticated, their capacity to distort democratic processes, manufacture consensus, and polarize communities intensifies [11]. Social media algorithms, optimized for engagement, systematically amplify misinformation and create epistemic bubbles—corroding civic discourse. Concurrently, biased training data perpetuates algorithmic exclusion, marginalizing underrepresented groups in critical domains (e.g., finance, healthcare). This requires a methodological shift: from reactive mitigation of harms after deployment to proactive embedding of equity and accountability into system design. AI must transition from reactive bias mitigation to proactive equity engineering, embedding justice as a core architectural principle rather than an add-on compliance feature [23].

5.1. Inclusivity as a Technical Imperative

Mitigating exclusion requires radical diversity in AI development ecosystems. Homogeneous teams—dominated by narrow demographic and disciplinary perspectives—inevitably encode blind spots into systems, as evidenced by facial recognition failures affecting darker-skinned women. Conversely, multidisciplinary teams (spanning ethicists, sociologists, and impacted communities) expose latent biases during design phases, not post-deployment. For example, participatory design frameworks enable marginalized groups to co-define fairness metrics, ensuring systems reflect pluralistic values. This is not merely ethical hygiene; it is technical necessity. Diverse teams improve model robustness by stress-testing edge cases and expanding solution spaces—directly enhancing predictive accuracy and user trust [14].

5.2. Environmental Sustainability – The Hidden Cost of Scale

AI's climate impact now constitutes an ethical crisis. Training a single large language model (e.g., GPT-3) emits aprox. 552 tonnes of CO—equivalent to 123 gasoline-powered vehicles driven for a year [30]. The 2024 OECD AI Principles explicitly mandate environmental due diligence, requiring [24]:

- Algorithmic efficiency: Sparse architectures, quantization, and federated learning to reduce computation.
- Infrastructure decarbonization: Migration to renewable-powered data centers.
- Resource circularity: Model reuse and hardware lifecycle management.
- Critically, sustainability cannot be outsourced to "greenwashing." The DATA Framework's Environmental Well-being pillar operationalizes this via carbon-aware training protocols and energy impact audits—aligning technical choices with planetary boundaries.

´

5.3. Governing Emergent Technologies – Generative AI and Autonomy

Generative AI and autonomous systems introduce unprecedented governance challenges [23]:

- Generative models enable synthetic media weaponization (deepfakes, disinformation) while obscuring provenance.

- Autonomous agents (e.g., self-driving vehicles) decentralize decision-making, complicating accountability.
- Current static regulations falter against these dynamics. Effective governance requires:
- Adaptive maturity models: Tiered compliance frameworks evolving with technical capabilities (e.g., requiring watermarking for generative tools at Stage 1, progressing to real-time provenance tracking at Stage 3).
- Human-AI co-regulation: "Shared autonomy" protocols where humans intervene in high-stakes edge cases (e.g., medical diagnostics).
- Liability mirrors: Legal structures assigning responsibility across developers, deployers, and users based on decision traceability.

5.4. The Case for Transnational Regulatory Harmonization

Fragmented national policies (e.g., EU AI Act vs. U.S. sectoral guidelines) create compliance chaos and enforcement gaps. International standards are essential to [20]:

- Prevent jurisdiction shopping: Developers migrating to lax regulatory regimes.
- Counter asymmetric threats: Deepfake-driven disinformation crossing borders.
- Ensure baseline protections: Universal requirements for explainability, impact assessments, and human oversight.
- The DATA Framework's Dynamic Governance component enables this through automated compliance mapping, translating regulations (GDPR, AI Act, OECD standards) into verifiable system requirements. This harmonizes oversight without stifling innovation.

6. AI System Readiness Assessment Framework: A Six-Dimensional Evaluation Model for Production Deployment

The preceding analysis establishes an urgent imperative: existing AI governance approaches fail to operationalize ethical principles into deployable safeguards, as evidenced by high-profile failures like COMPAS and regulatory gaps in frameworks like ALTAI. While Section 2 exposed critical trust deficits through criminological and bias analysis, and Section 3 revealed limitations in static compliance regimes, this chapter addresses the core research gap — the absence of actionable readiness quantification. We therefore introduce the Ethical AI Impact Matrix Framework, transforming abstract ethical requirements (Section 3) into a six-dimensional evaluative engine. By converting accountability, interpretability, equity, robustness, privacy, and reliability into quantifiable deployment thresholds, this model provides the missing link between ethical aspiration (EU Guidelines) and operational reality—enforcing trust through verifiable production readiness.

The transition from experimental AI prototypes to production-ready systems represents one of the most critical challenges in contemporary artificial intelligence development. While significant attention has been devoted to algorithmic performance optimization and technical validation, the systematic evaluation of AI system readiness for operational deployment remains an under-explored domain requiring structured methodological approaches. We propose a comprehensive six-dimensional readiness assessment framework specifically designed to evaluate AI systems' preparedness for production deployment. The framework addresses the fundamental question: "Is this AI system ready for real-world deployment?" by providing quantifiable metrics across accountability, interpretability, equity, robustness, privacy, and reliability domains. Through systematic evaluation of these dimensions, organizations can make informed decisions about system deployment timing, required improvements, and ongoing operational requirements. The framework employs a four-point scale (0-4) for each dimension, providing greater granularity than traditional three-point scales while maintaining practical usability. This expanded scale enables more nuanced differentiation between system capabilities and supports progressive improvement tracking throughout the development lifecycle.

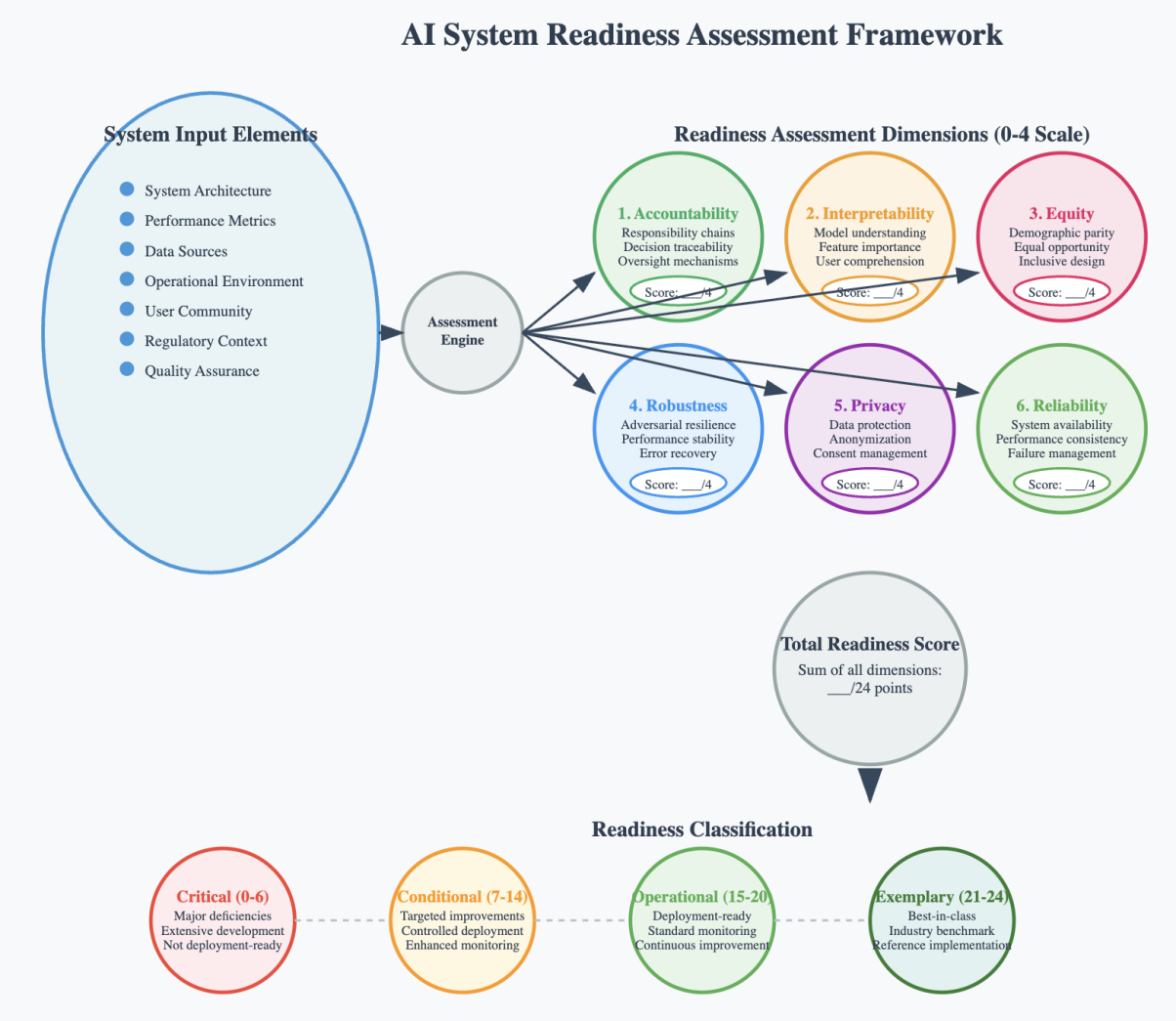

Figure 2: Ethical AI Impact Matrix for High-Stakes Decision Systems

Each of the six ethical requirements translates into concrete operational safeguards. For example, accountability requires defined audit trails and responsibility assignments for each decision; interpretability entails interfaces that allow end-users and auditors to contest predictions; equity requires subgroup performance audits to ensure calibration across demographics; robustness is operationalized through stress-testing on adversarial and out-of-distribution data; privacy requires data minimization and secure storage protocols; and reliability demands continuous monitoring of error rates in live deployments. In the case of COMPAS, failure to implement such safeguards (e.g., lack of transparency and subgroup audits) illustrates how ethical principles remain disconnected from deployment reality — precisely the gap EAIIM seeks to close.

6.1. Quantitative Assessment Framework

The EAIIM reflects the iterative and holistic nature of evaluation. The central assessment engine processes inputs from seven key system elements and distributes evaluation across six dimensional domains, creating a comprehensive readiness profile, figure 2.

It employs a standardized 0-4 point scale for each dimension, where 0 indicates inadequate preparation and 4 represents exemplary readiness. This five-level scale provides sufficient granularity for meaningful differentiation while maintaining practical usability. The total possible score of 24 points (6 dimensions × 4 points each) enables comprehensive readiness profiling and comparative analysis across different

$$S_{final} = \left(\sum_{i=0}^{6} s_i \right) X\ (1 - 0.4)^{n_{zeros}}\ X\ (1 - 0.1)^{n_{low}}$$

Figure 3: Formula 1 - EAIIM classification system

AI systems. Each scoring level represents distinct capability thresholds:

- **Level 0** (Inadequate): Fundamental requirements not met, significant deficiencies present
- **Level 1** (Basic): Minimal requirements met, substantial improvements needed
- **Level 2** (Developing): Moderate capabilities demonstrated, some enhancements required
- **Level 3** (Proficient): Strong capabilities demonstrated, minor improvements beneficial
- **Level 4** (Exemplary): Outstanding capabilities demonstrated, industry-leading implementation

6.2. Readiness Classification System

The framework establishes four primary readiness categories based on aggregate scores:

- **Critical (0-6 points)** systems demonstrate fundamental inadequacies that preclude safe deployment. These systems require extensive development work across multiple dimensions and pose unacceptable risks to stakeholders. Critical systems should not be deployed under any circumstances and require comprehensive redesign or substantial improvement before reconsideration.
- **Conditional (7-14 points)** systems show partial readiness but require targeted improvements and enhanced supervision for safe deployment. These systems may be suitable for limited pilot deployments with appropriate safeguards, continuous monitoring, and rapid improvement cycles. Conditional deployment should include specific mitigation measures addressing identified deficiencies.
- **Operational (15-20 points)** systems demonstrate comprehensive readiness for standard deployment with routine monitoring requirements. These systems have addressed major readiness factors and established appropriate governance structures. Operational systems require ongoing monitoring and continuous improvement but can be deployed with confidence in standard production environments.
- **Exemplary (21-24 points)** systems represent industry-leading implementations that exceed standard readiness requirements. These systems demonstrate outstanding capabilities across all dimensions and serve as benchmarks for other AI system development efforts. Exemplary systems may be suitable for high-stakes applications and can serve as reference implementations for organizational learning.

A formula is proposed figure 3, to invalidate that systems with classification of 0 or lower than 2 in certain dimensions still get a positive evaluation. It is required that each dimension has to meet a minimum threshold (e.g., Level 2) before deployment, ensuring that no critical domain such as interpretability or equity is neglected even if aggregate scores are high. With this formula, a 40% penalization occurs whenever a classification of 0 exists, and 10% penalization for low evaluations. this formula ensures that if at least one criterion has a zero score, then the system is classified as Critical High Risk, and it is impossible to compensate for a critical zero with maximum scores in other criteria.

6.3. Implementation Methodology

The framework supports integration into existing AI development lifecycles through staged assessment protocols. Initial assessments during system design enable early identification of readiness challenges and proactive improvement planning. Intermediate assessments during development provide progress tracking and course correction opportunities. Final assessments before deployment ensure comprehensive readiness verification and stakeholder confidence. Effective implementation requires systematic

engagement with diverse stakeholders throughout the assessment process. In practice, scoring should be conducted by a multidisciplinary panel including compliance auditors, internal QA teams, and domain experts, following standardized rubrics for each dimension. This ensures consistency and accountability in evaluation. Technical teams provide detailed system knowledge and capability assessment. Domain experts contribute application-specific requirements and risk evaluation. End users offer usability and acceptability perspectives. Regulatory representatives ensure compliance and governance adequacy. This multi-stakeholder approach ensures comprehensive evaluation and stakeholder buy-in.

The EAIIM incorporates ongoing assessment mechanisms that enable continuous readiness monitoring and improvement. Regular reassessment cycles ensure that systems maintain readiness levels as operational conditions change. Performance monitoring provides real-time feedback on system behavior and readiness indicators. Incident analysis enables learning and improvement cycle implementation.

Applied to COMPAS, EAIIM would score interpretability at Level 0–1 due to proprietary opacity, equity at Level 1–2 due to demonstrated racial disparities, and accountability at Level 1 given limited oversight mechanisms. These low values would place COMPAS in the 'Critical' category, illustrating how the framework flags unsuitability for deployment without significant redesign.

7. Conclusions

The rapid advancement of AI has brought transformative benefits across multiple domains but has also introduced significant ethical and societal challenges. This work has highlighted critical concerns related to algorithmic bias, transparency, and fairness, using COMPAS as a case study to illustrate the risks inherent in AI-driven decision-making. The evidence suggests that while AI systems can improve efficiency, they must be carefully designed and continuously monitored to prevent the reinforcement of systemic inequalities. The debate surrounding COMPAS underscores the broader issue of algorithmic opacity and the consequences of proprietary decision-making systems that lack public scrutiny. Addressing these challenges requires regulatory frameworks, such as the European Union's AI guidelines and the NIST AI Risk Management Framework, to ensure AI operates ethically and equitably. A new framework to mitigate the identified issues in this work was proposed.It provides a comprehensive, systematic approach to evaluating AI system preparedness for production deployment. By focusing on six critical dimensions of readiness, the framework enables organizations to make informed decisions about system deployment timing, required improvements, and ongoing operational requirements. The four-point scoring system enables nuanced assessment while maintaining practical usability for decision-making purposes. The four-tier classification system provides clear deployment guidance that balances innovation enablement with responsible AI practices. This framework represents a significant advancement in AI system evaluation methodology, moving beyond narrow technical metrics to encompass comprehensive readiness assessment. Its adoption has the potential to improve AI system quality, enhance stakeholder confidence, and support the responsible development of beneficial artificial intelligence applications. Finally, to achieve a responsible integration of AI into society, a collaborative effort among researchers, policymakers, and industry stakeholders is essential. Additionally, sustainability considerations must be incorporated into AI development to reduce its environmental footprint and ensure long-term social responsibility. Ultimately, the future of AI hinges on striking a balance between innovation and ethical responsibility. By prioritizing fairness, accountability, and human-centered AI design, intelligent systems that not only enhance decision-making but also contribute to a more just and equitable society can be built.

Acknowledgments

This work has been supported through the FCT project 2024.07420.IACDC, https://doi.org/10.54499/2024.07420.IACDC.

The work of Pedro Oliveira was supported by the doctoral Grant PRT/BD/154311 /2022 financed by the Portuguese Foundation for Science and Technology (FCT), and with funds from European Union,

under MIT Portugal Program.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias: risk assessments in criminal sentencing (2016), ProPublica https://www. propublica. org (2016).
[2] B. Green, Y. Chen, The accuracy, fairness, and limits of predicting recidivism, Science Advances 4 (2018) eaao5580. doi:10.1126/sciadv.aao5580.
[3] A. Khademi, V. Honavar, Algorithmic bias in recidivism prediction: A causal perspective, in: Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI), volume 34, 2020, pp. 3831–3838. doi:10.1609/aaai.v34i03.5855.
[4] R. Akers, W. Jennings, Social learning theory. the handbook of criminological theory, 230-240. west sussex, 2015.
[5] J. P. Williams, Subcultural theory: Traditions and concepts, Polity, 2011.
[6] T. Glad, L. Ljung, Control theory, CRC press, 2018.
[7] S. H. Manne, A communication theory of sociopathic personality, American Journal of Psychotherapy 21 (1967) 797–807.
[8] L. Hannon, Criminal opportunity theory and the relationship between poverty and property crime, Sociological Spectrum 22 (2002) 363–381.
[9] M. Yazdi, E. Zarei, S. Adumene, A. Beheshti, Navigating the power of artificial intelligence in risk management: a comparative analysis, Safety 10 (2024) 42.
[10] S. Thomas, The Fairness Fallacy: Northpointe and the COMPAS Recidivism Prediction Algorithm, Ph.D. thesis, Columbia University, 2023.
[11] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, et al., The malicious use of artificial intelligence: Forecasting, prevention, and mitigation, arXiv preprint arXiv:1802.07228 (2018).
[12] M. del Pilar Roa Avella, J. E. Sanabria-Moyano, Use of the compas algorithm in the criminal process and the risks to human rights, Rev. Brasileira de Direito Processual Penal 8 (2022) 275.
[13] Z. Zhu, Z. Jin, H. Hu, M. Xue, R. Sun, S. Camtepe, P. Gauravaram, H. Chen, Ai-compass: A comprehensive and effective multi-module testing tool for ai systems, arXiv preprint arXiv:2411.06146 (2024).
[14] A. L. Washington, How to argue with an algorithm: Lessons from the compas-propublica debate, Colo. Tech. LJ 17 (2018) 131.
[15] C. Novelli, M. Taddeo, L. Floridi, Accountability in artificial intelligence: What it is and how it works, Ai & Society 39 (2024) 1871–1882.
[16] R. Calegari, G. G. Castañé, M. Milano, B. O'Sullivan, et al., Assessing and enforcing fairness in the ai lifecycle, in: IJCAI, volume 2023, International Joint Conferences on Artificial Intelligence, 2023, pp. 6554–6562.
[17] M. Ducret, E. Wahal, D. Gruson, S. Amrani, R. Richert, M. Mouncif-Moungache, F. Schwendicke, Trustworthy artificial intelligence in dentistry: learnings from the eu ai act, Journal of Dental Research 103 (2024) 1051–1056.
[18] S. Stampernas, C. Lambrinoudakis, A framework for compliance with regulation (eu) 2024/1689 for small and medium-sized enterprises, Journal of Cybersecurity and Privacy 5 (2025) 40.
[19] A. Fedele, C. Punzi, S. Tramacere, et al., The altai checklist as a tool to assess ethical and legal implications for a trustworthy ai development in education, Computer Law & Security Review 53 (2024) 105986.

[20] N. Polemi, I. Praça, K. Kioskli, A. Bécue, Challenges and efforts in managing ai trustworthiness risks: a state of knowledge, Frontiers in big Data 7 (2024) 1381163.
[21] C. Lalonde, O. Boiral, Managing risks through iso 31000: A critical analysis, Risk management 14 (2012) 272–300.
[22] R. Dotan, B. Blili-Hamelin, R. Madhavan, J. Matthews, J. Scarpino, Evolving ai risk management: A maturity model based on the nist ai risk management framework, arXiv preprint arXiv:2401.15229 (2024).
[23] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, D. Mané, Concrete problems in ai safety, arXiv preprint arXiv:1606.06565 (2016).
[24] R. Verdecchia, J. Sallou, L. Cruz, A systematic review of green ai, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 13 (2023) e1507.