AI Act-Oriented Risk Assessment Framework for High-Risk Intelligent Systems

Jordan Butler
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
jordan.work@oregonstate.edu

Andreas D. Rose
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
contactandreas@ucf.edu

Sven Perez
Department of Computer Science, University of North Texas, Denton, TX, USA.
hellosven@unt.edu

Jeffrey A. Korhonen
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
jakorhonen@unh.edu

Abstract

The European Union's Artificial Intelligence Act (AI Act) establishes a tiered regulatory framework that imposes stringent requirements on high-risk AI systems, necessitating robust and transparent risk assessment methodologies. This paper presents a comprehensive system-level risk assessment framework specifically designed for high-risk intelligent systems as defined by the AI Act. The framework integrates technical, governance, and socio-technical dimensions to evaluate risks across the entire lifecycle of an AI system, from conception and training to deployment and continuous monitoring. We examine structural trade-offs between system performance and safety, architectural patterns that support compliance, and the infrastructure needed for ongoing validation. The treatment addresses issues of robustness, fairness, sustainability, and accountability, drawing on insights from large-scale systems engineering, regulatory science, and AI ethics. Through cross-domain comparisons and case illustrations, the framework is shown to be adaptable to sectors such as healthcare, critical infrastructure, criminal justice, and employment. The paper further explores policy implications and forward-looking perspectives on how such frameworks can evolve as AI capabilities and regulatory landscapes mature. This work aims to provide researchers, practitioners, and policymakers with a rigorous yet practical foundation for operationalizing AI Act requirements in the design and governance of high-risk intelligent systems.

Keywords

AI Act, risk assessment, high-risk AI systems, governance, system architecture, fairness, robustness, sustainability, socio-technical systems, regulatory compliance.

1. Introduction

The rapid proliferation of artificial intelligence (AI) across critical domains has triggered a global regulatory response aimed at curbing potential harms without stifling innovation. The

European Union's AI Act [1] represents the most comprehensive legislative effort to date, introducing a risk-based classification that subjects high-risk AI systems to mandatory conformity assessments, transparency obligations, human oversight, and continuous monitoring. Existing risk assessment frameworks, such as those originating from software engineering [2] or general product safety [3], are insufficient to capture the unique challenges posed by AI systems that learn, adapt, and operate under uncertainty. Moreover, the AI Act itself prescribes high-level requirements but does not define an operational risk assessment methodology. This gap motivates the development of an AI Act-oriented risk assessment framework that translates regulatory mandates into actionable, system-level design and governance practices.

The framework proposed in this paper adopts a socio-technical perspective, recognizing that high-risk intelligent systems are not merely technical artifacts but are embedded in organizational, legal, and societal contexts. Structural trade-offs between competing objectives—such as predictive accuracy and fairness, or innovation speed and safety—must be examined from the earliest design stages. The framework emphasizes architectural decisions, infrastructure for continuous validation, and governance mechanisms that ensure accountability across the system's lifespan. By integrating concepts from large-scale systems engineering [4], responsible AI [5], and regulatory compliance, the framework provides a systematic approach for organizations to demonstrate conformity with AI Act requirements while maintaining system performance and sustainability.

2. Background on the EU AI Act and High-Risk Classification

The AI Act [6] categorizes AI systems into four risk levels: unacceptable, high, limited, and minimal. High-risk systems include those used in biometric identification, critical infrastructure management, educational and vocational training, employment, access to essential services, law enforcement, migration and border control, and administration of justice [7]. For each high-risk system, the Act requires a risk management process, training data governance, technical documentation, record-keeping, transparency, human oversight, and accuracy and robustness standards [1]. The regulation also mandates that high-risk systems undergo a conformity assessment, which for many systems involves third-party review by notified bodies.

A key challenge lies in operationalizing these requirements. While the Act provides high-level principles, it does not specify how risk should be measured, what threshold triggers high-risk classification beyond the list, or how to assess systemic risks arising from the interaction of multiple AI systems [8]. Furthermore, the concept of risk in the AI Act is influenced by product safety legislation, where risk is defined as the combination of probability and severity of harm [9]. However, AI systems introduce novel sources of harm, such as discriminatory bias, privacy intrusion, and loss of human autonomy, which are difficult to quantify using traditional probabilistic methods.

Several scholarly works have analyzed the AI Act's implications. One study [5] examines the tension between regulatory certainty and technological agility, while others [10] critique the lack of clear technical metrics for conformity. The framework proposed here addresses these gaps by providing a structured methodology that can be tailored to different high-risk domains while maintaining rigor and interpretability for auditors.

3. A Proposed Risk Assessment Framework

The AI Act-Oriented Risk Assessment Framework (AORA-F) is organized into four interconnected phases: contextual analysis, system architecture review, operational risk evaluation, and continuous oversight. Each phase operates at the system level, considering not only the AI model but also the data pipeline, deployment environment, human-AI interaction points, and governance structure.

The first phase, contextual analysis, involves mapping the intended purpose, domain regulations, stakeholder landscape, and potential sources of harm. This phase aligns with the AI Act's requirement for an intended purpose statement and risk classification [6]. For example, a credit scoring system used for loan approvals falls under high-risk employment and essential services, and must be analyzed in light of existing financial regulations and fair lending laws. The analysis identifies relevant legal baselines, ethical guidelines, and organizational policies that impose constraints on design decisions.

The second phase, system architecture review, examines technical design choices that influence risk. This includes data acquisition and preprocessing, model selection, training procedures, validation protocols, explainability mechanisms, and human oversight interfaces. The framework emphasizes modular architectures that allow independent verification of components [11]. For instance, a separate bias detection module that can be validated against fairness metrics without retraining the entire model reduces the risk of hidden discriminatory patterns. Architectural decisions also affect robustness: adversarial training or input sanitization can mitigate certain failure modes, but may degrade accuracy or increase computational cost, representing a classic trade-off that must be documented and justified.

The third phase, operational risk evaluation, quantifies and qualifies risks based on evidence from testing, simulation, and real-world monitoring. Rather than assigning a single risk score, the framework produces a multi-dimensional risk profile covering functional safety, fairness, privacy, transparency, and sustainability. These dimensions are assessed using both quantitative metrics (e.g., demographic parity difference, robust accuracy under distribution shift) and qualitative judgments from expert review panels. The evaluation is dynamic: as the system is monitored post-deployment, risk estimates are updated using feedback loops [12].

The fourth phase, continuous oversight, establishes governance processes for ongoing risk management. This includes a risk register, incident response protocols, periodic reassessment schedules, and audit trails. The framework mandates that governance bodies include diverse stakeholders, such as domain experts, end-user representatives, and ethics officers, to guard against groupthink and capture multiple perspectives [13]. Oversight mechanisms must be embedded into the system architecture itself, for example through automated logging of all decisions involving high-risk outcomes.

4. Structural Trade-offs and Architectural Considerations

Every high-risk AI system embodies a set of design trade-offs that directly affect its regulatory compliance and societal impact. One prominent trade-off is between model performance and fairness. Many fairness metrics, such as equalized odds or predictive parity, conflict with accuracy objectives when base rates differ across groups [14]. The framework requires that such trade-offs be explicitly documented and justified in the risk management file, with evidence that alternative designs were explored and that any disparity is minimized to the extent technically feasible.

Another structural trade-off involves transparency versus intellectual property protection. The AI Act mandates technical documentation and explainability, but proprietary algorithms or

training data may be considered trade secrets. The framework encourages a layered transparency approach: public-facing summaries that describe system behavior in non-technical language, detailed technical reports shared with regulators under confidentiality agreements, and open-sourcing of bias audit tools. This architecture respects both regulatory requirements and commercial interests, as argued in prior work on accountability in algorithmic systems [15].

Sustainability also presents a trade-off. High-risk systems, especially deep learning models, consume significant energy during training and inference. The AI Act does not explicitly address environmental impact, but the broader European Green Deal and corporate sustainability goals demand that organizations consider carbon footprint. The framework incorporates energy efficiency as a risk dimension, with metrics such as total energy per prediction or per training run. Organizations may need to choose between a more accurate but energy-intensive model and a less accurate but greener alternative. This trade-off must be weighed against potential harm: in critical healthcare applications, a small increase in accuracy may save lives, justifying higher energy consumption, whereas in low-stakes applications, greener models may be preferred.

Architectural patterns that facilitate compliance include microservice-based decomposition, where each component (e.g., data ingestion, model inference, output filtering) can be independently validated and replaced. Such architecture also supports human-in-the-loop decision making: for high-risk outputs, the system can route decisions to a human reviewer before final action is taken. The framework specifies minimum requirements for human oversight, including the frequency of review, the quality of information presented to the human, and the ability for the human to override the system [16].

5. Governance and Policy Implications

Governance structures are central to the effectiveness of any risk assessment framework. The AI Act requires that high-risk systems have a conformity assessment procedure, but does not prescribe internal governance arrangements. The framework proposes a three-tier governance model: an operational risk committee at the project level, a corporate AI ethics board overseeing multiple systems, and an external audit function. Each tier has distinct responsibilities, from daily monitoring to strategic direction.

Policy implications extend beyond individual organizations. The framework must be interoperable with national competent authorities and notified bodies that perform conformity assessments. Standardized reporting formats and risk metric definitions are needed to ensure consistency across sectors [17]. The framework supports this by defining a common ontology of risk categories and evidence types, enabling regulators to compare systems and accumulate knowledge.

Another policy dimension concerns the liability of upstream actors, such as data providers and model developers. The AI Act holds deployers primarily accountable, but the framework encourages a shared responsibility model where each entity in the value chain signs risk management agreements [18]. This aligns with the concept of socio-technical accountability, where legal and technical responsibilities are clearly allocated.

The framework also addresses the need for continuous learning in regulation. As AI systems evolve, so must risk assessments. The AI Act envisions regular reviews, but the framework goes further by mandating that any significant modification to the system triggers a partial

reassessment. This prevents so-called "regulatory drift" where systems gradually change their behavior post-market without re-evaluation [19].

6. Deployment, Sustainability, and Robustness

Deployment context dramatically influences risk. An AI system used in a hospital radiology department faces different failure modes than the same system used in telemedicine. The framework requires a deployment-specific risk analysis, considering environmental factors such as network reliability, user expertise, and potential for adversarial manipulation. For example, a medical diagnostic tool deployed in a low-bandwidth rural clinic may need to operate offline and with lower-resolution images, increasing the risk of misdiagnosis. The framework flags such scenarios and suggests mitigation strategies, including edge computing with local fallback.

Sustainability is treated not only as energy efficiency but also as longevity of the system. High-risk systems often require periodic retraining as data distributions change. The framework mandates a data governance plan that specifies retraining triggers, version control, and impact assessments for each new model release. This ensures that the system remains robust over time without incurring unsustainable maintenance costs.

Robustness in the AI Act context refers to the system's ability to perform under expected and outlier conditions. The framework incorporates stress testing, adversarial validation, and sensitivity analysis. For high-risk systems, robustness must be demonstrated not only for the model but for the entire pipeline, including pre-processing and post-processing steps. A notable challenge is that robustness to one type of disturbance (e.g., Gaussian noise) may come at the cost of robustness to another (e.g., structural perturbations). The framework requires a comprehensive robustness profile covering multiple perturbation types and a rationale for accepted trade-offs [20].

7. Fairness and Accountability

Fairness in high-risk AI systems is a multi-faceted construct. The AI Act prohibits unfair discrimination but does not mandate a specific fairness metric. The framework adopts a pluralistic approach, requiring that deployers assess fairness from several normative perspectives: individual fairness (similar cases treated similarly), group fairness (demographic parity, equal opportunity), and procedural fairness (transparency of decision processes). This aligns with recommendations from algorithmic fairness research [21]. The evaluation must be intersectional, considering how multiple protected attributes (e.g., race and gender) interact to produce compounded disparities.

Accountability mechanisms are built into the framework's oversight phase. Every high-stakes decision must be logged with sufficient metadata to allow ex-post review. The log should include the input data, model output, confidence scores, human override if any, and the identity of the human reviewer. These logs form the basis for periodic audits and for investigating incidents. The framework also specifies that accountability extends to the system's designers: an algorithm impact assessment must be published before deployment, similar to data protection impact assessments under GDPR [22].

8. Case Illustrations and Cross-Domain Comparisons

To demonstrate the framework's applicability, we consider three high-risk domains: healthcare diagnostics, employment screening, and predictive policing. In healthcare, an AI system for detecting cancer from radiological images must balance sensitivity and specificity.

The framework's contextual analysis would identify the high cost of false negatives (missed cancers) versus false positives (unnecessary biopsies). An architectural recommendation might be to use a two-stage system: a sensitive initial screening followed by a specific confirmatory model. The operational risk evaluation would include calibration metrics and robustness to different scanner brands. Continuous oversight would mandate regular performance monitoring across demographic groups to detect drift.

In employment screening, an AI resume parser used by an employer must avoid bias based on gender or ethnicity. The framework would require fairness audits using historical hiring data, and would recommend a human-in-the-loop for key decisions such as interview invitations. A structural trade-off arises between using sophisticated natural language processing that may encode subtle biases versus simpler keyword matching that may miss qualified minority candidates. The framework demands transparent documentation of the chosen approach.

Predictive policing systems raise profound ethical and legal questions. The framework's contextual analysis would highlight the risk of reinforcing systemic biases if historical arrest data encode discriminatory practices. Architectural recommendations include limiting the system's outputs to resource allocation recommendations rather than direct targeting, and requiring a civilian oversight board. The operational risk evaluation would incorporate disparate impact simulation and community feedback mechanisms.

Cross-domain comparisons reveal common patterns: the need for transparent documentation, human oversight mechanisms, and continuous monitoring. However, the specific risk metrics and governance structures differ significantly, underscoring the framework's adaptability rather than prescriptive uniformity.

9. Forward-Looking Perspectives

As AI technology advances, the risk assessment framework must evolve. Emerging paradigms such as foundation models and generative AI introduce systemic risks that are not yet well captured by the AI Act's high-risk categories [23]. The framework's modular structure allows it to incorporate new risk dimensions, such as amplification of misinformation or model collapse due to synthetic data. Additionally, the framework could be extended to cover collective risks arising from multiple interacting AI systems, a topic of growing concern in multi-agent settings and AI swarms.

Regulatory harmonization is another future challenge. The AI Act may become a global benchmark, but different jurisdictions (e.g., the US, China, Japan) have different regulatory philosophies. The framework should be designed with a core set of requirements that are internationally recognized, while allowing for local adaptations. International standards bodies, such as ISO and IEEE, are developing guidelines that could be mapped onto the framework's phases [24].

Finally, the framework must remain practical for small and medium-sized enterprises (SMEs) that may lack resources for extensive risk assessment. The research community should develop lightweight toolkits and automated audit tools that operationalize parts of the framework without overwhelming developers. Balancing comprehensiveness with feasibility is an ongoing design challenge.

10. Conclusion

This paper has presented a comprehensive risk assessment framework for high-risk intelligent systems in the context of the EU AI Act. The framework integrates technical architecture,

governance, and socio-technical considerations across four phases: contextual analysis, system architecture review, operational risk evaluation, and continuous oversight. It addresses structural trade-offs, sustainability, robustness, fairness, and accountability, and provides a basis for regulatory compliance and responsible innovation. Through case illustrations and cross-domain comparisons, the framework demonstrates flexibility while maintaining rigor. As AI systems become more pervasive and powerful, such frameworks will be essential for ensuring that their benefits are realized without undermining safety, equity, or trust. Future work should focus on empirical validation, tooling development, and alignment with emerging international norms.

References

1. European Commission. (2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence (AI Act). COM(2021) 206 final.
2. IEEE. (2019). IEEE Standard for Software Quality Assurance Processes (IEEE Std 730-2014).
3. International Organization for Standardization. (2018). Risk management — Guidelines (ISO 31000:2018).
4. Crawley, E., Cameron, B., & Selva, D. (2015). System architecture: Strategy and product development for complex systems. Pearson.
5. Floridi, L., & Cowls, 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
6. European Parliament. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. Official Journal of the European Union.
7. Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the draft EU Artificial Intelligence Act. Computer Law Review International, 22(4), 97-112.
8. Smuha, N. A. (2021). From a 'race to the bottom' to a 'race to the top'?: The European Union's approach to the regulation of artificial intelligence. European Journal of Risk Regulation, 12(1), 1-17.
9. European Commission. (2001). Directive 2001/95/EC on general product safety.
10. Ebers, M. (2021). Regulating AI in the EU: A new regulatory framework for artificial intelligence. Journal of Law and the Biosciences, 8(1), lsab005.
11. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31-38.
12. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.
13. Schiff, D., Biddle, J., Borenstein, J., & Laas, K. (2021). What's next for AI ethics, policy, and governance? A global overview. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 153-158.
14. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29.

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

16. Green, B., & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-24.

17. Kaminski, M. E., & Rubinfeld, D. L. (2023). The regulation of AI in the European Union: A primer. Journal of European Competition Law & Practice, 14(2), 95-105.

18. Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. International Data Privacy Law, 7(2), 76-99.

19. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of the Conference on Fairness, Accountability, and Transparency, 59-68.

20. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations.

21. Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning: Limitations and opportunities. MIT Press.

22. Edwards, L., & Veale, M. (2017). Slave to the algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for. Duke Law & Technology Review, 16(1), 18-84.

23. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

24. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019). Ethically aligned design (2nd ed.). IEEE.

25. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399.