

NAVIGATING **AI** **COMPLIANCE**

PART 2: RISK MITIGATION STRATEGIES FOR SAFEGUARDING AGAINST FUTURE FAILURES

MARIAMI TKESHELASHVILI
TIFFANY SAADE

MARCH 2025



IST

Institute for
SECURITY + TECHNOLOGY

Navigating AI Compliance, Part 2
Risk Mitigation Strategies for Safeguarding Against Future Failures

March 2025
Author: Mariami Tkeshelashvili and Tiffany Saade
Design: Sophia Mauro

The Institute for Security and Technology and the authors of this report invite free use of the information within for educational purposes, requiring only that the reproduced material clearly cite the full source.

Copyright 2025, The Institute for Security and Technology
Printed in the United States of America

About the Institute for Security and Technology

Uniting technology and policy leaders to create actionable solutions to emerging security challenges

Technology has the potential to unlock greater knowledge, enhance our collective capabilities, and create new opportunities for growth and innovation. However, insecure, negligent, or exploitative technological advancements can threaten global security and stability. Anticipating these issues and guiding the development of trustworthy technology is essential to preserve what we all value.

The Institute for Security and Technology (IST), the 501(c)(3) critical action think tank, stands at the forefront of this imperative, uniting policymakers, technology experts, and industry leaders to identify and translate discourse into impact. We take collaborative action to advance national security and global stability through technology built on trust, guiding businesses and governments with hands-on expertise, in-depth analysis, and a global network.

We work across three analytical pillars: the **Future of Digital Security**, examining the systemic security risks of societal dependence on digital technologies; **Geopolitics of Technology**, anticipating the positive and negative security effects of emerging, disruptive technologies on the international balance of power, within states, and between governments and industries; and **Innovation and Catastrophic Risk**, providing deep technical and analytical expertise on technology-derived existential threats to society.

Learn more: <https://securityandtechnology.org/>

Acknowledgments

This work is inherently collaborative. As researchers, conveners, and facilitators, the Institute for Security and Technology (IST) is immensely grateful to the members of the AI Risk Reduction multi stakeholder working group for their insights, dedication, willingness to engage in honest and healthy debate, and the time that each of them generously volunteered to this effort.

We are also immensely grateful for the generous support of Patrick J. McGovern Foundation, whose funding allowed us to continue this project through the lens of IST's Applied Trust & Safety program.

AI is too vast a set of tools, capabilities, and communities for any one organization to manage the risks and opportunities on its own. This effort reflects the cross-sectoral, public-private efforts needed more broadly across the ecosystem to ensure AI is beneficial for us all. We extend our gratitude to the following experts who contributed to this paper by providing their feedback, guidance, and participation in the multi stakeholder meetings:

- | | |
|---------------------|-------------------------|
| » Chloe Autio | » Brian Judge |
| » Henriette Cramer | » Monica Lopez |
| » Matthew da Mota | » Alexander Reese |
| » Hadassah Drukarch | » Alyssa Lefavre Škopac |
| » Avijit Ghosh | » Justin Sherman |
| » Kamya Jagadish | » Peter Slattery |
| » Katherine Johnson | » Akash Wasil |

Finally, the authors extend their gratitude to Steve Kelly and Philip Reiner for the support and strategic guidance they provided in the drafting and refining of this report.

Contents

- Executive Summary1**
- Recap of *Navigating AI Compliance, Part 1* 3**
- Introduction 4**
- Methodology 5**
 - The AI Lifecycle Stages7
- Return on Investment for Implementing Strong Compliance Practices 8**
- Risk Mitigation Strategies for Safeguarding Against Future Failures 10**
 - Data Collection and Preprocessing11
 - Model Training and Evaluation 14
 - Model Application 17
 - User Interaction 18
 - Ongoing Monitoring and Maintenance 19
- Conclusion 20**

Executive Summary

Historical trends do not wholly dictate the future of AI. While the first installment of this report acknowledged the importance of historic lessons, we can make deliberate choices to shape what comes next. We hope that “Navigating AI Compliance, Part 2: Safeguarding Against Future Failures” will guide decision-makers in fostering societal trust in AI systems, all while preventing the repetition of past mistakes.

This report, the second in a two-part series, presents 39 risk mitigation strategies for avoiding institutional, procedural, and performance failures of AI systems (see [Risk Mitigation Strategies for Safeguarding Against Future Failures](#)). These strategies aim to enhance user trust in AI systems and maximize product utilization. AI builders and users, including AI labs, enterprises deploying AI systems, as well as state and local governments, can use and implement a selection of the 22 technical and 17 policy-oriented risk mitigation strategies presented in this report according to their needs and risk thresholds.

Through implementing these practices, organizations building and utilizing AI systems not only reduce regulatory risk exposure and build user trust for their product, but they could also attract top talent, gain a competitive edge, enhance their financial performance, and increase the lifetime value of their solutions. Based on our research and the results of stakeholder engagement, we emphasize to AI builders and users the following nine recommendations from the complete list of 39:

- » **Implement proportional compliance measures for high-impact AI applications.** AI builders and users should consider which compliance measures are most appropriate for their work, especially when building or deploying AI systems in sensitive or high-impact areas. This consideration should be proportional to factors such as the intended use, potential risks, and application domain—ranging from entertainment to critical sectors like national security, healthcare, and finance.
- » **Acknowledge and address acceptable risks in AI development and deployment.** Unintended consequences are not to be confused with compliance failure. Still, these unplanned effects should be acknowledged by developers, builders, and regulators as they consider thresholds of acceptable tolerance for the enhanced risks associated with exposed attack surfaces and features or functionalities of AI that are not yet thoroughly understood or anticipated.

- » **Prioritize data management and privacy practices to maintain user trust.** Implementing proper data management and privacy-enhancing practices will protect user rights, maintain trust, and comply with data protection regulations. Measures such as privacy-preserving technologies, content provenance features, and user consent mechanisms can alleviate procedural failures.
- » **Implement robust cybersecurity controls for AI infrastructure protection and enhanced reliability.** Cybersecurity controls, red-teaming, fail-safe mechanisms, and other techniques protect AI systems from attacks and strengthen their reliability in various scenarios. Security guardrails may alleviate or preempt both performance and procedural failures.
- » **Utilize safety and risk assessments to proactively mitigate AI harms.** Safety and risk assessment procedures, such as incident reporting frameworks and AI safety benchmarks at different stages of the lifecycle, identify and mitigate possible harms before they occur—potentially mitigating both procedural and performance failures.
- » **Design and implement compliance and AI literacy training for staff.** Training should be mandatory for all staff members involved in the AI supply chain, from data providers to model developers and deployers. All staff members utilizing AI tools in some manner should also obtain a minimum set of AI literacy skills through the training.
- » **Build trust by implementing transparency mechanisms.** Transparency and interpretability mechanisms such as model cards, data cards, and disclosure frameworks are necessary to build user and stakeholder trust, facilitate accountability, and enable informed decision-making.
- » **Enhance AI explainability and disclosure frameworks to improve understanding of system behavior.** Efforts to increase the explainability of AI systems, supplemented with disclosure frameworks for model evaluation, allow both builders and users to better understand the behavior patterns and outputs of these systems and potentially safeguard against performance failures.
- » **Employ strategies for non-discriminatory AI.** Bias mitigation strategies across model training, data collection, and ongoing monitoring and maintenance, in addition to adversarial debiasing, can prevent performance failures and help to ensure fairness while preventing discriminatory outcomes in AI systems.

Recap of *Navigating AI Compliance, Part 1: Tracing Failure Patterns in History*

History often rhymes with and echoes through the present and future. Through this lens, the first installment of this two-part report series examined past compliance failures across various industries—from nuclear power to financial services—as a source of definitions, frameworks, and lessons learned to help AI builders and users navigate today’s complex compliance landscape.¹ Our analysis of eleven case studies from AI-adjacent industries revealed three distinct categories of failure:

- » **Institutional failures stem from a lack of executive commitment to create a culture of compliance, establish necessary policies, or empower success through the organizational structure, leading to foreseeable failures.**
- » **Procedural failures are the result of a misalignment between an institution’s established policies and its internal procedures and staff training required to adhere to those policies.**
- » **Performance failures result when employees fail to follow an established process, or an automated system fails to perform as intended, leading to an undesirable result.**

By studying failures across sectors, we uncovered critical lessons about risk assessment, safety protocols, and oversight mechanisms that can guide AI innovators in this era of rapid development. One of the most prominent risks is the tendency to prioritize rapid innovation and market dominance over safety. The case studies demonstrated a crucial need for transparency, robust third-party evaluation and verification, and comprehensive data governance practices, among other safety and security measures.

Though today’s AI regulatory landscape remains fragmented, we identified five main sources of AI governance—laws and regulations, guidance, norms, standards, and organizational policies—to provide AI builders and users with a clear direction for the safe, secure, and responsible development of AI. Therefore, we defined “compliance failure” within the AI

¹ Mariami Tkeshelashvili and Tiffany Saade, “Navigating AI Compliance, Part 1: Tracing Failure Patterns in History,” Institute for Security and Technology, December 2024, <https://securityandtechnology.org/wp-content/uploads/2024/12/Navigating-AI-Compliance.pdf>.

ecosystem as the failure to align with and adhere to any of these governance mechanisms, whether publicly announced or confidential.

Part 1 of this report series concluded by addressing AI's unique compliance issues stemming from its ongoing evolution and complexity. Ambiguous AI safety definitions and the rapid pace of development challenge efforts to govern it—including AI's adoption within regulated industries—while interpretability challenges hinder the development of compliance mechanisms. Furthermore, the rapid advent of agentic AI will introduce added complexity and blur the lines of liability in an increasingly automated world.

Introduction

As illustrated in the first of this two-part report, any technology can fail and cause harm. But failure of a technology product that has achieved ubiquity in the marketplace can generate magnified effects—which is the essence of concentration risk. As AI quickly trends toward ubiquity, the risks of AI system failures and their ripple effects are further magnified by a trend toward AI autonomy. It is therefore all the more important to manage these risks and alleviate, pre-empt, and avoid future failures.

How exactly can AI builders and users defend against future failure risks? What are the benefits of proactively implementing compliance practices? This report aims to:

- » Provide AI builders with technical and policy-oriented risk mitigation strategies for avoiding compliance failures in the future. AI builders are defined in Part 1 of this report series as “individuals or organizations responsible for developing the models including AI labs, startups, and tech companies.”²
- » Provide AI users with technical and policy-oriented risk mitigation strategies for responsible deployment of AI systems. AI users are defined in Part 1 as “all other entities who deploy or utilize the technology, including enterprises integrating AI systems into their services and internal operations.”³
- » Illuminate the various ways in which sound compliance practices can generate return on investment (ROI).

This report's proposed risk mitigation strategies are inspired by lessons learned from past compliance failures noted in Part 1 and co-created by the working group members listed in the

2 Mariami Tkeshelashvili and Tiffany Saade, “Navigating AI Compliance, Part 1.”

3 Mariami Tkeshelashvili and Tiffany Saade, “Navigating AI Compliance, Part 1.”

Methodology

In order to integrate existing AI governance frameworks into our thinking, we analyzed a set of AI norms, standards, and regulations—both binding and voluntary—to distill the main themes and patterns for technical and policy mitigation strategies across the AI lifecycle. The sources we integrated are: voluntary commitments such as the Hiroshima Process;⁵ the Organization for Economic Co-operation and Development’s (OECD’s) AI Framework;⁶ United Kingdom AI Framework;⁷ work of the Coalition for Content Provenance and Authenticity (C2PA);⁸ National Institute of Standards and Technology (NIST) AI Risk Management Framework;⁹ ISO/IEC standard 42001;¹⁰ and binding regulatory frameworks such as the European Union’s AI Act¹¹ and General Data Protection Regulation (GDPR).¹² Additionally, we integrated relevant

12 “Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation),” *Official Journal of the European Union* 119/1 (May 4, 2016), <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679>.

elements from IEEE standards 7000¹³ and 7002,¹⁴ and specific ISO standards which are not exclusive to AI systems but establish important standards for ethical system design and data privacy.

The risk mitigation strategies presented in this report both leverage and are aligned to IST's previously articulated AI Lifecycle Framework, which breaks down the complex process of AI development into manageable stages.¹⁵ This structured approach ensures a comprehensive understanding of each phase, making it easier to develop and implement specific risk mitigation strategies.

13 Institute of Electrical and Electronics Engineers (IEEE), "IEEE Standard Model Process for Addressing Ethical Concerns during System Design," IEEE 7000-2021, September 15, 2021, <https://standards.ieee.org/ieee/7000/6781/>.

14 Institute of Electrical and Electronics Engineers (IEEE), "IEEE Standard for Data Privacy Process," IEEE 7002-2022, April 19, 2022, <https://standards.ieee.org/ieee/7002/6898/>.

15 Louie Kangeter, "A Lifecycle Approach to AI Risk Reduction: Tackling the Risk of Malicious Use Amid Implications of Openness," Institute for Security and Technology, June 2024, <https://securityandtechnology.org/wp-content/uploads/2024/06/A-Lifecycle-Approach-to-AI-Risk-Reduction.pdf>.

The AI Lifecycle Stages

The AI Lifecycle Framework breaks down the complex process of AI development into manageable stages. This structured approach ensures a comprehensive understanding of each phase, making it easier to target specific risk mitigation strategies effectively.

Data Collection & Preprocessing

Gathering raw data from various sources and preparing it for use in training the AI model; cleaning the data, handling missing values, normalizing data formats, and augmenting datasets to enhance their quality and relevance.

Model Architecture

Designing and structuring the AI model, including the choice of algorithms, network topology (for neural networks), and other foundational elements that define how the model processes data and makes predictions.

Model Training & Evaluation

Using preprocessed data to teach the AI model to recognize patterns and make predictions; testing the model's performance using separate validation datasets to ensure it generalizes well to new, unseen data.

Model Deployment

Integrating the trained AI model into a production environment where it can be accessed and used by end-users or other systems. Includes setting up the necessary infrastructure, such as servers and APIs, to support model operation.

User Interaction

Designing the interfaces and interactions through which end-users engage with AI-powered applications. This includes user experience (UX) design, user interface (UI) design, and accessibility considerations.

Model Application

Developing applications that use the deployed AI models to perform specific tasks. Includes designing the software and systems that leverage the AI model's capabilities for various user needs.

Ongoing Monitoring and Maintenance

Continuously tracking the performance of deployed AI models and applications, addressing any issues that arise, and updating models as needed to ensure they remain effective and secure.

Return on Investment for Implementing Strong Compliance Practices

Non-compliance in building and deploying AI systems can result in consequences such as reputational harm, erosion of public trust, and fines.^{16,17,18,19,20} Instead of being reactive, AI builders and users can adopt proactive compliance practices that help accelerate and amplify the value both builders and users can derive from the technology.²¹ There are various ways in which strong compliance practices can generate ROI:

- » **Reduced regulatory risk exposure.** Given the rapid proliferation of AI tools, industries utilizing these technologies are expected to face increasing scrutiny from regulators. Proactively implementing safety, security, privacy, transparency, and anti-bias measures—and a compliance program to oversee their implementation—can help prevent unexpected and costly harms, their associated litigation, and reputational implications. For instance, in December 2024, just four compliance fines totaled up to a hefty quarter billion euros for failing to comply with GDPR.²² Both GDPR and the EU AI Act have extraterritorial reach, which means that some of the provisions apply to companies that are not physically based in the EU but offer products and services within the EU market. For instance, an AI lab based in the United States which makes their AI tools available to EU users is subject to GDPR, EU AI Act, and other regulations governing the EU market.

16 Elvira Pollina and Alvis Armellini, “Italy Fines OpenAI 15 Million Euros over Privacy Rules Breach,” *Reuters*, December 20, 2024, <https://www.reuters.com/technology/italy-fines-openai-15-million-euros-over-privacy-rules-breach-2024-12-20/>.

17 Nikitha Anand, “The High Cost of Non-Compliance: Penalties Issued for AI under Existing Laws,” *Holistic AI*, March 28, 2024, <https://www.holisticai.com/blog/high-cost-non-compliance-penalties-under-ai-law>.

18 Natasha Lomas, “MWC’s Organizer Slapped with GDPR Fine over Biometrics ID Checks Due Diligence,” *TechCrunch*, May 8, 2023, <https://techcrunch.com/2023/05/08/gsma-mwc-aedp-gdpr-dpia-fine/>.

19 David Shepardson, “Lingo Telecom Agrees to \$1 Million Fine over AI-Generated Biden Robocalls,” *Reuters*, August 21, 2024, <https://www.reuters.com/technology/artificial-intelligence/lingo-telecom-agrees-1-million-fine-over-ai-generated-biden-robocalls-2024-08-21/>.

20 CMS.Law, “GDPR Enforcement Tracker - List of GDPR Fines,” last accessed February 2025, <https://www.enforcementtracker.com/?insights>.

21 Velu Sinha, Julie Coffman, Richard Fleming, Bill Groves, and Maria Teresa Tejada, “Adapting Your Organization for Responsible AI,” *Bain*, January 2, 2024, <https://www.bain.com/insights/adapting-your-organization-for-responsible-ai/>.

22 CMS.Law, “GDPR Enforcement Tracker - List of GDPR Fines,” last accessed February 2025, <https://www.enforcementtracker.com/?insights>.

- » **Competitive advantage.** Strong compliance practices provide a competitive advantage for both AI system builders and the enterprises adopting the systems. A recent report from Bain reveals that organizations with an effective approach to responsible AI doubled their profit impact from their AI efforts compared to those organizations that lack such an approach.²³
- » **Access to government procurement-shaped markets.** The U.S. government's procurement policies and preferences make and shape markets. In 2023 alone, the U.S. government invested more than \$100 billion in information technology products and services.²⁴ A company that complies with the relevant standards in AI space will be better prepared to compete in government procurement-shaped markets. For example, logging features required to be turned on by default in government procurement-shaped markets as a result of Executive Order 14028 on Improving the Nation's Cybersecurity then became industry standard for all cloud services in the market. Additionally, as the AI market becomes one of the largest and one of the most valued in the geopolitical and economic race to the top, governments will likely increase their investments into the development of frontier models, likely favoring those companies that have more robust security standards in place.²⁵
- » **Ability to recruit and retain talent.** Based on the working group members' experiences and observations, organizations that prioritize responsible AI development and deployment practices have an edge in attracting top talent who increasingly seek workplaces committed to responsible innovation. A strong ethical framework enhances employee morale and loyalty, fostering an environment where skilled professionals want to contribute and grow. This talent pipeline is crucial for both model capability development as well as scaling AI products into new markets worldwide.
- » **Increased lifetime value.** By investing in responsible AI practices, companies can build stronger relationships with customers, partners, and employees, leading to higher satisfaction and loyalty. For customers, this translates to increased lifetime value to the company, as satisfied customers are more likely to return and advocate for the brand, ultimately boosting long-term profitability. Proactively addressing AI compliance concerns can safeguard an organization's reputation over time. Companies that navigate these challenges effectively are better positioned to withstand scrutiny and maintain public trust, helping their brands remain resilient against potential controversies.

23 Velu Sinha et al., "Adapting Your Organization for Responsible AI."

24 The White House, "Fact Sheet: OMB Issues Guidance to Advance the Responsible Acquisition of AI in Government," press release, October 2, 2024, <https://bidenwhitehouse.archives.gov/omb/briefing-room/2024/10/03/fact-sheet-omb-issues-guidance-to-advance-the-responsible-acquisition-of-ai-in-government/>.

25 The White House, "Fact Sheet: OMB Issues Guidance."

- » **Investor appeal.** Enterprises that demonstrate compliance, particularly in emerging technologies like AI, are likely to attract more investment, as stakeholders increasingly consider security risks. A rigorous compliance program can indicate to investors that the company has a lower risk threshold, prompting new investors, and sustaining existing investors.^{26,27}

Risk Mitigation Strategies for Safeguarding Against Future Failures

It is important to note that no risk management strategy, or combination of strategies, will completely eliminate the possibility of an undesired outcome in any context. This is particularly true in AI, as bad actors aggressively test the limits of their capabilities and as the potential for “capability overhang”—defined as AI capabilities and aptitudes that were not envisioned by their developers but emerge nonetheless—increases.²⁸ As a result, it becomes challenging to preempt unforeseen risks arising from novel capabilities and to foresee how malicious actors could exploit them. As a rapidly developing frontier, AI has a limited track record from which to design and implement effective controls. It also follows that an unintended consequence should not always be attributed to a compliance failure, as not all AI risks and failure modes are yet well understood. Such negative experiences can instead serve as learning trials for builders, users, and regulators alike as they refine the state of practice in AI risk management.

The following table contains a selection of 22 technical and 17 policy-oriented risk mitigation strategies co-created by the working group members and other contributors for alleviating, pre-empting, or avoiding the three categories of compliance failure risks in the AI ecosystem.

26 Matthew White, Justin Daniels, and Javier Becerra, “AI Disclosures under the Spotlight: SEC Expectations for Year-End Filings,” Baker Donelson, January 10, 2025, <https://www.bakerdonelson.com/ai-disclosures-under-the-spotlight-sec-expectations-for-year-end-filings>.

27 Christopher Barlow, Brett Fleisher, David Simon, Nicola Kerr-Shaw, Melissa Muse, and Taylor Votek, “Rising Investment in AI Requires Financial Sponsors to Address Unique Risks,” Skadden, Arps, Slate, Meagher & Flom LLP, January 14, 2025, <https://www.skadden.com/insights/publications/2025/01/2025-insights-sections/the-deal-landscape/rising-investment-in-ai-requires-financial-sponsors>.

28 Zoë Brammer, “How Does Access Impact Risk?: Assessing AI Foundation Model Risk Along a Gradient of Access,” Institute for Security and Technology, December 2023, <https://securityandtechnology.org/wp-content/uploads/2023/12/How-Does-Access-Impact-Risk-Assessing-AI-Foundation-Model-Risk-Along-A-Gradient-of-Access-Dec-2023.pdf>.



Institutional failures

Lack of executive commitment to create a culture of compliance, establish necessary policies, or empower success through the organizational structure (e.g., risk and audit board committees, compliance officer role, quality assurance program), leading to foreseeable failures.



Procedural failures

Misalignments between an institution's established policies as compared to its internal procedures and staff training required to adhere to those policies.



Performance failures

An employee's failure to follow an established process, or an automated system's failure to perform as intended, leading to an undesirable result.

We recognize that implementing all 39 of the below strategies may not always be feasible. However, AI builders and users should consider which measures are appropriate according to their context. This consideration should be proportional to factors such as the intended use, potential risks, and application domain, which can range from entertainment and arts to national security, healthcare, and finance.

Data Collection and Preprocessing (for builders)		Types of risks mitigated
policy	1. Data collection requirements Ensure that the collection, processing, and maintenance of personal or other protected data takes place in accordance with a valid legal basis. For instance, ensure that explicit consent is obtained from individuals whose data is collected, with mechanisms to withdraw consent at any point.	 
technical	2. Privacy-preserving technologies Protect sensitive data during the training stage by implementing privacy-preserving technologies—such as differential privacy and homomorphic encryption—during data pre-processing so that, for example, the model does not learn personally identifiable information. Implement data encryption both at rest and in transit to prevent label flipping attacks and insecure data storage.²⁹	

²⁹ Databricks, “Databricks AI Security Framework (DASF),” 2024, <https://www.databricks.com/resources/whitepaper/databricks-ai-security-framework-dasf>.

technical	3. Data source transparency through data cards For each model, publish a “data card” that documents the model’s data sources, privacy measures, and preprocessing steps taken by its developers during the data collection and model training phases. ^{30,31,32,33}	
technical	4. Bias detection tools for dataset auditing Utilize automated bias detection tools to sift through training datasets and look for potential imbalances in attributes such as race, language, age, heritage, gender, viewpoint, etc. Ensure that the training data is tested for accuracy and truthfulness to avoid negatively influencing the model with non-factual information. Implement methods such as data augmentation or re-weighting to mitigate potential biases. ^{34,35}	

Model Architecture (for builders)		Types of risks mitigated
policy	5. Cross-functional AI compliance team Establish a cross-functional AI compliance team with representation from relevant corporate functions such as legal, product, engineering, data infrastructure, cybersecurity, ethics, and internal audit functions. The team can blend together organizational strategies at different lifecycle stages, harmonize internal policies and practices, and address emerging issues related to compliance. <i>(Note, this mitigation applies to this and all subsequent lifecycle phases.)</i>	
policy	6. Security program Design or implement existing, reliable, robust cybersecurity and physical security controls to secure model architecture and the infrastructure hosting the AI systems. Limit access to the system components to authorized personnel, with relevant aspects carefully managed, controlled, and monitored. <i>(Note, this mitigation applies to this and all subsequent lifecycle phases.)</i>	

- 30 Nathalie Baracaldo and Hayim Shaul, “Fully Homomorphic Encryption,” IBM Research, February 9, 2021, <https://research.ibm.com/topics/fully-homomorphic-encryption>.
- 31 Mahima Pushkarna, Andrew Zaldivar, Dan Nanas et al., “Data Cards Playbook,” People + AI Research, Google, March 5, 2021, <https://sites.research.google/datacardsplaybook/>. According to Google’s “Data Card Playbook,” data cards are “structured summaries of essential facts about various aspects of ML datasets needed by stakeholders across a project’s lifecycle for responsible AI development.”
- 32 Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson, “Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI,” *arXiv*, April 3, 2022, <https://doi.org/10.48550/arXiv.2204.01075>.
- 33 “Regulation (EU) 2024/1689 EU Artificial Intelligence Act,” *Official Journal of the European Union* 2024/1689 (July 7, 2024), <http://data.europa.eu/eli/reg/2024/1689/oj>.
- 34 Agnieszka Mikołajczyk-Bareła, Maria Ferlin, and Michał Grochowski, “Targeted Data Augmentation for Bias Mitigation,” *arXiv*, August 22, 2023, <https://arxiv.org/abs/2308.11386>.
- 35 “Pledge for a Trustworthy AI in the World of Work,” proceedings in the Summit for Action on Artificial Intelligence, Paris, February 11, 2025, <https://www.elysee.fr/emmanuel-macron/2025/02/11/pledge-for-a-trustworthy-ai-in-the-world-of-work>.

technical	7. Explainability by design Document and report an AI model’s features that explain its outputs, including the contribution of specific training data points, while integrating explainability frameworks that simplify complex machine learning models into easily understandable representations.^{36,37,38}	
technical	8. Threat model-informed design requirements Simulate a variety of adversarial attacks to test and improve the robustness of the model against malicious inputs to safeguard AI systems, especially in high-risk applications.³⁹	
technical	9. Anomaly detection Incorporate anomaly detection and continuous monitoring mechanisms into model architecture to identify unusual or malicious activity in real time and provide alerts for potential misuse.⁴⁰	
technical	10. Model cards Create a model card for each user-facing model that documents its architecture, including but not limited to performance metrics, explainability, safety measures, and robustness tests performed.^{41,42,43} Model cards can include documentation of the system’s intent, precise scope (i.e., intended use cases and known limitations), as well as any “out of scope” uses (i.e., what the model should not be used for) and the model’s known technical mitigations. Update model cards periodically with newly observed model performance metrics, including potential risks.	

36 European Parliament, “EU AI Act: First Regulation on Artificial Intelligence,” European Parliament, June 8, 2023, <https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>.

37 C3.AI, “LIME: Local Interpretable Model-Agnostic Explanations,” last accessed February 2025, <https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/>.

38 Arize AI, “Explainability in Machine Learning: Top Techniques,” Arize Machine Learning Course, January 11, 2024, <https://arize.com/blog-course/explainability-techniques-shap/>.

39 Jonas Rauber and Roland S. Zimmermann, “Welcome to Foolbox Native — Foolbox 3.3.3 Documentation,” Foolbox, 2021, <https://foolbox.readthedocs.io/en/stable/>.

40 Louie Kangeter, “A Lifecycle Approach to AI Risk Reduction.”

41 Margaret Mitchell et al., “Model Cards for Model Reporting,” *arXiv*, January 14, 2019, <https://arxiv.org/abs/1810.03993>. According to Google’s Model Cards Paper introduced in 2018, model cards “are short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic groups and intersectional groups that are relevant to the intended application domains. Model cards also disclose the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information.”

42 OECD.AI, “OECD Framework for the Classification of AI Systems.”

43 NIST, “NIST AI RMF Playbook,” July 8, 2022, <https://www.nist.gov/itl/ai-risk-management-framework/nist-ai-rmf-playbook>.

Model Training and Evaluation (for builders)		Types of risks mitigated
policy	11. AI safety benchmarks Task AI compliance teams to start establishing mandatory safety benchmarks for exceptionally capable models that stand to impact individuals and society in a highly contextual fashion (e.g., contextualize based on use within specific industries or affecting vulnerable population groups).^{44,45,46} Integrate AI safety benchmarks, and specify that models must pass certain safety criteria before deployment. Evaluate the models across multiple axes, such as accuracy, fairness, bias, and robustness—akin to safety certifications found in other industries (e.g., automotive crash tests).^{47,48} Specify that evaluations must be conducted on diverse datasets to mitigate risks of overfitting when models are deployed.⁴⁹	
policy	12. Benchmark hazard categories Benchmark hazard categories (e.g. hate speech, Child Sexual Abuse Material (CSAM), violence, drugs, etc.) to guide training data selection and prompt generation. Create labeled data with these safety categories in mind to improve how models classify and identify risks.⁵⁰	
policy	13. Model evaluation guidelines Craft model evaluation guidelines to include metrics around algorithmic transparency, which would require the documentation of all training datasets, algorithm choices, hyperparameter tuning, and metrics used to assess performance. These model evaluations should be repeated periodically during training, especially for models that learn continuously or adapt in real-time.^{51,52,53,54,55}	

44 European Parliament, “EU AI Act: First Regulation on Artificial Intelligence.”

45 Anthropic, “Anthropic’s Responsible Scaling Policy,” September 19, 2023, <https://www.anthropic.com/news/anthropics-responsible-scaling-policy>.

46 Anca Dragan, Helen King, and Allan Dafoe, “Introducing the Frontier Safety Framework,” Google DeepMind, December 17, 2024, <https://deepmind.google/discover/blog/introducing-the-frontier-safety-framework>.

47 MLCommons AI Safety Working Group, “Announcing a Benchmark to Improve AI Safety,” IEEE Spectrum, April 16, 2024, <https://spectrum.ieee.org/ai-safety-benchmark>.

48 Insurance Institute for Highway Safety (IIHS) and Highway Loss Data Institute (HLDI), “Vehicle Ratings,” last accessed January 24, 2025, <https://www.iihs.org/ratings>.

49 For example, after every epoch, developers can run safety benchmarks on the version of the model at hand to pinpoint emerging safety deficiencies, and these benchmarking results could serve as data points to potentially adjust training objectives (e.g., reinforce guardrails that help models avoid generating harmful responses).

50 “Pledge for a Trustworthy AI in the World of Work,” proceedings in the Summit for Action on Artificial Intelligence.”

51 International Standards Organization (ISO), “ISO/IEC DIS 42001,” 2023, <https://www.iso.org/standard/81230.html>.

52 NIST, “AI Risk Management Framework.”

53 OECD, “OECD Framework for the Classification of AI Systems,” *OECD Publishing*, no. 323 (February 2022), https://www.oecd.org/content/dam/oecd/en/publications/reports/2022/02/oecd-framework-for-the-classification-of-ai-systems_336a8b57/cb6d9eca-en.pdf.

54 IEEE Standards Association, “IEEE Standard Model Process for Addressing Ethical Concerns during System Design,” IEEE 7000-2021, September 15, 2021, <https://standards.ieee.org/ieee/7000/6781/>.

55 This component mirrors the UK framework’s mention of an iterative approach to risk management to address new risks as they come into shape.

technical	14. Data overfitting mitigations Guard against data overfitting, wherein a model performs well on training data but fails with new, unseen prompts. ⁵⁶ Use out-of-distribution data to ensure models generalize well to new prompts, rather than just performing well on benchmark-specific scenarios. ^{57,58}	
technical	15. Data provenance and watermarking Incorporate content provenance features in all model outputs, such as watermarks or metadata that can verify the origin and integrity of generated content. This can prevent bad actors from manipulating or misusing the model for harmful purposes and enhance traceability. ^{59,60}	
technical	16. Bug bounty programs Create bug bounty programs to incentivize others to identify and report previously unknown weaknesses in an AI model. ⁶¹	
technical	17. Privacy-preserving technologies Ensure AI systems are privacy compliant by integrating privacy-preserving technologies to minimize the danger of data exposure. This mitigation, and its performance, should be included in any privacy compliance reports. ⁶²	
technical	18. Bias monitoring and data integrity checks Monitor potential biases during training through techniques such as adversarial debiasing. Consider benchmarking model datasets on common fairness metrics such as demographic parity and equalized odds to mitigate bias. ^{63,64,65}	

⁵⁶ IBM, “What Is Overfitting?” 2024, <https://www.ibm.com/topics/overfitting>.

⁵⁷ To avoid data overfitting, developers can assess and compare the AI model’s performance on training versus test data, and track a number of potential discrepancies in performance between both. If the performance of the model is significantly better on the training data, overfitting may be the cause. Developers can also create an Out-Of-Distribution dataset, which comprises examples of data points not included in the training set to measure the model’s performance on the OOD dataset and compare it with the model’s performance on its regular training set to assess the model’s ability to generalize.

⁵⁸ Alexandre Bonnet, “What Is Out-of-Distribution (OOD) Detection?” Encord, September 15, 2023, <https://encord.com/blog/what-is-out-of-distribution-ood-detection/>.

⁵⁹ Coalition for Content Provenance and Authenticity, “Guiding Principles,” 2024, <https://c2pa.org/principles/>.

⁶⁰ Kizuna, “The Hiroshima AI Process: Leading the Global Challenge to Shape Inclusive Governance for Generative AI,” The Government of Japan, February 9, 2024, https://www.japan.go.jp/kizuna/2024/02/hiroshima_ai_process.html.

⁶¹ Louie Kangeter, “A Lifecycle Approach to AI Risk Reduction.”

⁶² The White House, “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,” October 30, 2023, <https://bidenwhitehouse.archives.gov/briefing-room/presidential-actions/2025/01/14/executive-order-on-advancing-united-states-leadership-in-artificial-intelligence-infrastructure/>.

⁶³ IBM, “AI Fairness 360,” aif360.res.ibm.com, accessed January 24, 2025, <https://aif360.res.ibm.com/>.

⁶⁴ Jenny Yang et al., “An Adversarial Training Framework for Mitigating Algorithmic Biases in Clinical Machine Learning,” *Npj Digital Medicine* 6, no. 1 (March 29, 2023), <https://doi.org/10.1038/s41746-023-00805-y>.

⁶⁵ IGI Global, “What Is Equalized Odds,” 2023, <https://www.igi-global.com/dictionary/fairness-challenges-in-artificial-intelligence/115386>.

technical	19. Secure training pipelines Train models in a secure environment with version control and cryptographic measures to prevent unauthorized changes to datasets or model parameters. Apply penetration testing on AI training environments to identify and address vulnerabilities that could be exploited for malicious purposes. Record model performance and evaluation metrics in model cards for future auditing.	
Model Deployment (for builders and users)		Types of risks mitigated
policy	20. Incident reporting and disclosure framework Develop an incident reporting and response framework that requires AI system breaches and incidents to be documented and tracked. Include steps to escalate and report violations, such as jailbreaking.^{66,67} This framework could be leveraged in periodic or ad hoc reporting to an organization's compliance team, particularly for new tools being developed or deployed.⁶⁸	
policy	21. Staff training Design and implement mandatory compliance training for staff members involved in the AI supply chain. Training modules should be role-specific and take into account geographic jurisdiction and use context. All staff members utilizing AI tools should also demonstrate minimum literacy of AI system functions and limitations, intended use, and potential impact.^{69,70} <i>(Note, this mitigation applies to this and all subsequent lifecycle phases.)</i>	
policy	22. Deployment plan Consider deploying the system following a predefined, approved plan that outlines the AI system's inventory, maintenance, roles of involved actors, timeline, and a context-specific testing and feedback strategy aligned with the model's risk profile. The plan should also account for resource issues such as memory, compute, network, storage, redundancy, and load balancing. It should define risk thresholds, and incorporate digital, physical, and environmental security procedures to safeguard system assets.⁷¹	

66 Thorn and All Tech Is Human, "Safety by Design for Generative AI: Preventing Child Sexual Abuse," Thorn Repository, 2024, <https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf>.

67 Zeqiu Wu et al., "Fine-Grained Human Feedback Gives Better Rewards for Language Model Training," *arXiv*, October 30, 2023, <https://doi.org/10.48550/arXiv.2306.01693>.

68 Sean McGregor et al., "To Err Is AI: A Case Study Informing LLM Flaw Reporting Practices," *arXiv*, October 15, 2024, <https://arxiv.org/pdf/2410.12104>.

69 European Commission, "First Rules of the Artificial Intelligence Act Are Now Applicable," Shaping Europe's Digital Future, 2025, <https://digital-strategy.ec.europa.eu/en/news/first-rules-artificial-intelligence-act-are-now-applicable>.

70 Oliver Yaros et al., "EU AI Act: Ban on Certain AI Practices and Requirements for AI Literacy Come into Effect," Mayer Brown LLP, January 31, 2025, <https://www.mayerbrown.com/en/insights/publications/2025/01/eu-ai-act-ban-on-certain-ai-practices-and-requirements-for-ai-literacy-come-into-effect>.

71 UK Government, "National AI Strategy."

technical	23. Transparency measures Document and publicize (as appropriate) comparisons of a new AI model with existing models, infrastructure and tools, data accessibility, accuracy, interpretability, complexity, training time, and scalability. Implement model disclosure frameworks that include extended model cards, automated verification with reproducible testing and validation mechanisms, an adjudication process to fairly assess models, as well as a dynamic scope for models to adapt to emerging common uses.^{72,73}	
technical	24. System integration Integrate AI models into existing technical architectures and legacy systems to ensure they are practicable, accessible, and user-centric on both the back-end and front-end. Ensure that system integration processes account for compatibility with legacy systems, potential performance degradation, and potential data integration challenges.⁷⁴ Consider first testing in a sandbox to discover compatibility issues prior to integration.	

Model Application (for builders and users)		Types of risks mitigated
policy	25. Application-specific security controls When designing or deploying a specific AI tool, consider creating a decision tree to help choose which AI tool to deploy.⁷⁵ The decision tree should differ for AI tools used internally versus those used for business-to-user or business-to-business interactions.	
technical	26. Query rate limits Set a limit on the number of queries a user can input into an AI model within a specific timeframe to mitigate AI model abuse, including through automated means.^{76,77}	
technical	27. Human in the loop Mandate the inclusion of human oversight and control mechanisms in AI applications, especially for high-risk or sensitive use cases, to prevent fully autonomous unsanctioned actions. Define specific use cases in which agentic AI capabilities will provide operational advantages (e.g., increase productivity or efficiency) and cases in which keeping the human in the loop is essential for taking specific actions. Implement appropriate human-feedback loops and checks to assess the AI decision-making process and intervene when needed.	

⁷² Sven Cattell, Avijit Gosh, and Lucie-Aimée Kaffee, “View of Coordinated Flaw Disclosure for AI: Beyond Security Vulnerabilities,” *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, no. 1 (2024), <https://doi.org/10.1609/aies.v7i1.31635>.

⁷³ Anthropic, “Anthropic’s Transparency Hub,” last updated February 27, 2025, <https://www.anthropic.com/transparency>.

⁷⁴ NIST, “AI Risk Management Framework.”

⁷⁵ U.S. Department of Energy, “Cybersecurity Considerations for Procurement,” Federal Energy Management Program, October 2024, <https://www.energy.gov/femp/cybersecurity-considerations-procurement>.

⁷⁶ OpenAI, “OpenAI O1 and O1-Mini Usage Limits on ChatGPT and the API,” 2025, <https://help.openai.com/en/articles/9824962-openai-o1-preview-and-o1-mini-usage-limits-on-chatgpt-and-the-api>.

⁷⁷ Anthropic, “Rate Limits,” last accessed February 2025, <https://docs.anthropic.com/en/api/rate-limits>.

User Interaction (for builders and users)		Types of risks mitigated
policy	28. User consent Develop policies to ensure users are informed prior to an AI system making a decision on their behalf. For systems supporting high-impact use cases such as employment, financial, or healthcare decisions, provide users with clear explanations (using model cards or other techniques) of how decisions are made and how to appeal them. Ensure that user-AI interactions are governed by clear user consent mechanisms.⁷⁸	
policy	29. Robust user feedback loops Integrate mechanisms for users to provide feedback or contest decisions made by the AI system, to protect user autonomy and promote ethical engagement.⁷⁹	
policy	30. Education programs for end-users Implement programs to educate end-users about the limitations and proper use of an AI model, including safety measures to consider while interacting with the model. This would potentially increase public trust in AI by promoting informed interactions.	
technical	31. “Opt out” option for end-users Provide an explicit option for users to ‘opt out’ of processes in which decisions are made automatically by AI models and provide the option for human operators to be involved instead. Ensure that users are notified when an AI system is involved in generating content, advice, decisions, or actions and are provided with clear explanations of the criteria behind these outcomes.^{80,81}	
technical	32. Watermarking techniques Adopt watermarking techniques to identify AI-generated outputs for users’ awareness.⁸² While watermarking is not a holistic solution and can be vulnerable to tampering, it is a preliminary step to help users distinguish between traditionally produced and AI-generated content.	

⁷⁸ International Standards Organization (ISO), “ISO/IEC DIS 42001,” ISO, 2023, <https://www.iso.org/standard/81230.html>.

⁷⁹ International Standards Organization (ISO), “ISO/IEC DIS 42001.”

⁸⁰ “Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation),” *Official Journal of the European Union* 119/1 (May 4, 2016), <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679>.

⁸¹ Institute of Electrical and Electronics Engineers (IEEE), “IEEE Standard Model Process for Addressing Ethical Concerns during System Design,” IEEE 7000-2021, September 15, 2021, <https://standards.ieee.org/ieee/7000/6781/>.

⁸² Restack, “Watermarking Techniques in AI” Restack.io, 2025, <https://www.restack.io/p/ai-in-iot-answer-watermarking-techniques-cat-ai>.

Ongoing Monitoring and Maintenance (for builders and users)		Types of risks mitigated
policy	33. AI compliance reviews Task the AI Compliance Team to conduct periodic reviews during which models are audited to ensure continued alignment with relevant regulations, frameworks, and internal policies. Document and update all audits in the model cards to maintain transparency.^{83,84,85,86}	
policy	34. Responsible information sharing Uphold clear processes for responsibly sharing AI safety and security information with relevant stakeholders (i.e., governments, industry, civil society), to include security risks, potential vulnerabilities, and ways to mitigate misuse.⁸⁷	
policy	35. System transition and decommission Ensure that the AI system adheres to a transition or decommissioning plan that complies with applicable laws and regulations, protecting users' privacy and data rights, disposing of sensitive materials, and retaining system documentation for developers and the organization.	
policy	36. Third-party reviews Integrate periodic independent reviews to assess an AI model against safety, security, and performance quality metrics. These reviews could also include pre-deployment risk assessments and can be informed by insights from AI governance and policy-focused organizations.⁸⁸	
technical	37. Monitoring for model drift Use automated monitoring systems to track model performance over time and detect model drift or data drift. Implement mechanisms that can be triggered in the event a model starts behaving unpredictably, which might lead to humans retraining it.	

83 National Institute of Standards and Technology, “AI Risk Management Framework.”

84 “Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation),” *Official Journal of the European Union* 119/1 (May 4, 2016), <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679>.

85 European Parliament, “EU AI Act: First Regulation on Artificial Intelligence,” European Parliament, June 8, 2023, <https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>.

86 Reto Grubenmann and Flavia Masoni, “ISO/IEC 42001: The Latest AI Management System Standard,” KPMG, accessed January 26, 2025, <https://kpmg.com/ch/en/insights/technology/artificial-intelligence-iso-iec-42001.html>.

87 G7, “Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems.”

88 Monika Viktorova and Hadassah Drukarch, “Operationalizing Independent Review in AI Governance,” Responsible AI, November 25, 2024, <https://www.responsible.ai/operationalizing-independent-review-in-ai-governance/>.

technical	38. Model termination guidelines Develop clear emergency response protocols that specify under what circumstances an AI system would immediately be shut down, how this process would be carried out, and how it can be verified.	
technical	39. Monitoring protocols and logging Ensure that AI systems are designed to log all operational activities and AI-generated outputs such as reports, predictions, recommendations, and to provide the relevant stakeholders access to the recorded information. ^{89,90}	

Conclusion

Charting a path towards effective AI compliance measures requires the coordinated efforts of diverse stakeholders throughout the AI ecosystem. While this paper offers actionable risk mitigation strategies that AI builders and users can implement, there remains a need for broader collaboration on AI compliance. Safeguarding against future failures in the AI ecosystem requires a multidisciplinary approach; a technology sector that has a potential and ambition for universality should take insights from a broader array of stakeholders, including philosophers, ethicists, anthropologists, linguists, psychologists, and practitioners in human-computer interaction, user experience, and other disciplines.

AI ecosystem stakeholders should accelerate information sharing among verified researchers from leading labs, universities, and startups to diffuse best practices and methods for responsible AI development. By doing so, AI ecosystem stakeholders will have access to continuous learning resources, and AI builders and users will not have to choose between innovating and being responsible.

⁸⁹ European Parliament, “EU AI Act: First Regulation on Artificial Intelligence.”

⁹⁰ NIST, “AI Risk Management Framework.”



INSTITUTE FOR SECURITY AND TECHNOLOGY

www.securityandtechnology.org

info@securityandtechnology.org

Copyright 2025, The Institute for Security and Technology