AI Governance Handbook

By Khullani M. Abdullahi, J.D.

Introduction

Artificial Intelligence (AI) is transforming industries and society, but its rapid adoption raises concerns about ethics, safety, and compliance. AI Governance refers to the frameworks and processes ensuring AI is developed and used responsibly and in alignment with laws and values. This primer provides a comprehensive overview of AI Governance, AI Safety, Trustworthy AI, Responsible AI practices, and risk management. It is structured by audience – offering tailored insights for AI practitioners, compliance officers, executives, and policymakers – to address their specific concerns.

This primer will dig into key legal and regulatory frameworks (such as ISO/IEC 23894 and 42001, the EU AI Act, NIST AI RMF, GDPR, CCPA, etc.), privacy and security standards, and technical aspects of AI safety. We also provide a glossary clarifying core concepts and present six AI Maturity Stages frameworks (each with seven stages) for Governance, Safety, Trust & Transparency, Responsible AI, Risk Management, and Compliance. Visual frameworks and best practices at each maturity stage are included to help organizations benchmark and improve their AI governance efforts.

Legal, Policy, and Regulatory Frameworks for AI Governance

AI systems must comply with an evolving landscape of laws, regulations, and standards designed to address their unique risks. Key frameworks include international standards (ISO/IEC), national and regional laws (like the EU AI Act), and industry guidelines. Below we analyze some of the most relevant governance frameworks:

ISO/IEC 42001 (AI Management System Standard)

Published in December 2023, ISO/IEC 42001 is the first global standard for AI management systems[1]. It provides a certifiable framework for organizations to establish and continuously improve their AI governance processes[1]. This standard is akin to ISO 9001 (quality management) or ISO 27001 (information security), but tailored to AI. ISO 42001 focuses on ethics, transparency, accountability, bias mitigation, safety, and privacy – covering the essential elements of trustworthy AI development and deployment[1]. By implementing ISO 42001, organizations create an internal governance system ensuring AI projects are managed responsibly. The standard calls for defining an AI governance policy, senior leadership commitment, risk management processes, resource allocation for AI oversight, and operational controls for responsible AI throughout the AI lifecycle[1]. Importantly, ISO 42001 is sector-agnostic and for organizations of all sizes, providing a holistic approach to manage AI-related risks and opportunities across an entire organization[1]. Achieving ISO 42001 compliance can also demonstrate to regulators and customers that an organization adheres to recognized AI governance best practices.

ISO/IEC 23894 (AI Risk Management)

ISO/IEC 23894 is a companion standard providing detailed guidance on managing AI risks[1]. It essentially adapts the generic risk management principles of ISO 31000 to the AI context[2]. ISO 23894 guides organizations in integrating AI risk assessment into their processes – identifying risks across the AI lifecycle (from data collection and model training to deployment), evaluating their severity, and treating them with appropriate controls[1]. For

example, it covers processes to detect and mitigate bias or security vulnerabilities in AI models. Together, ISO 23894 (risk management) and ISO 42001 (management system) form a coherent toolkit: one establishes the governance structure, and the other provides risk-specific procedures. These ISO standards are international and voluntary, but they are likely to become baselines for compliance as regulators and customers increasingly expect organizations to follow them for AI governance consistency[1].

EU AI Act

The European Union's AI Act officially entered into force on August 1, 2024, marking a historic milestone in AI regulation. The Act is being implemented in phases, with key compliance deadlines extending through 2027. This legislation adopts a risk-based approach to AI governance, categorizing systems into four risk tiers with corresponding obligations[23]:

- **Unacceptable risk:** Banned applications, such as government-led social scoring, manipulative AI targeting vulnerable groups, and real-time biometric identification in public spaces.
- **High risk:** Critical systems, including those used in healthcare, recruitment, and law enforcement, subject to stringent oversight and conformity assessments.
- **Limited risk:** Systems requiring transparency measures, such as notifying users when interacting with AI chatbots.
- **Minimal risk:** Low-impact AI systems with minimal regulatory requirements.

The phased implementation includes several critical dates. As of February 2, 2025, the use of banned AI systems must cease. By August 2, 2025, provisions related to general-purpose AI models and penalties will take effect. High-risk AI system obligations will be enforceable starting August 2, 2026, with additional provisions extending into 2027. Non-compliance carries steep penalties: up to €35 million or 7% of global annual turnover for prohibited uses and €15 million or 3% for breaches of high-risk requirements[24]. Smaller businesses face scaled-down fines proportional to their turnover.

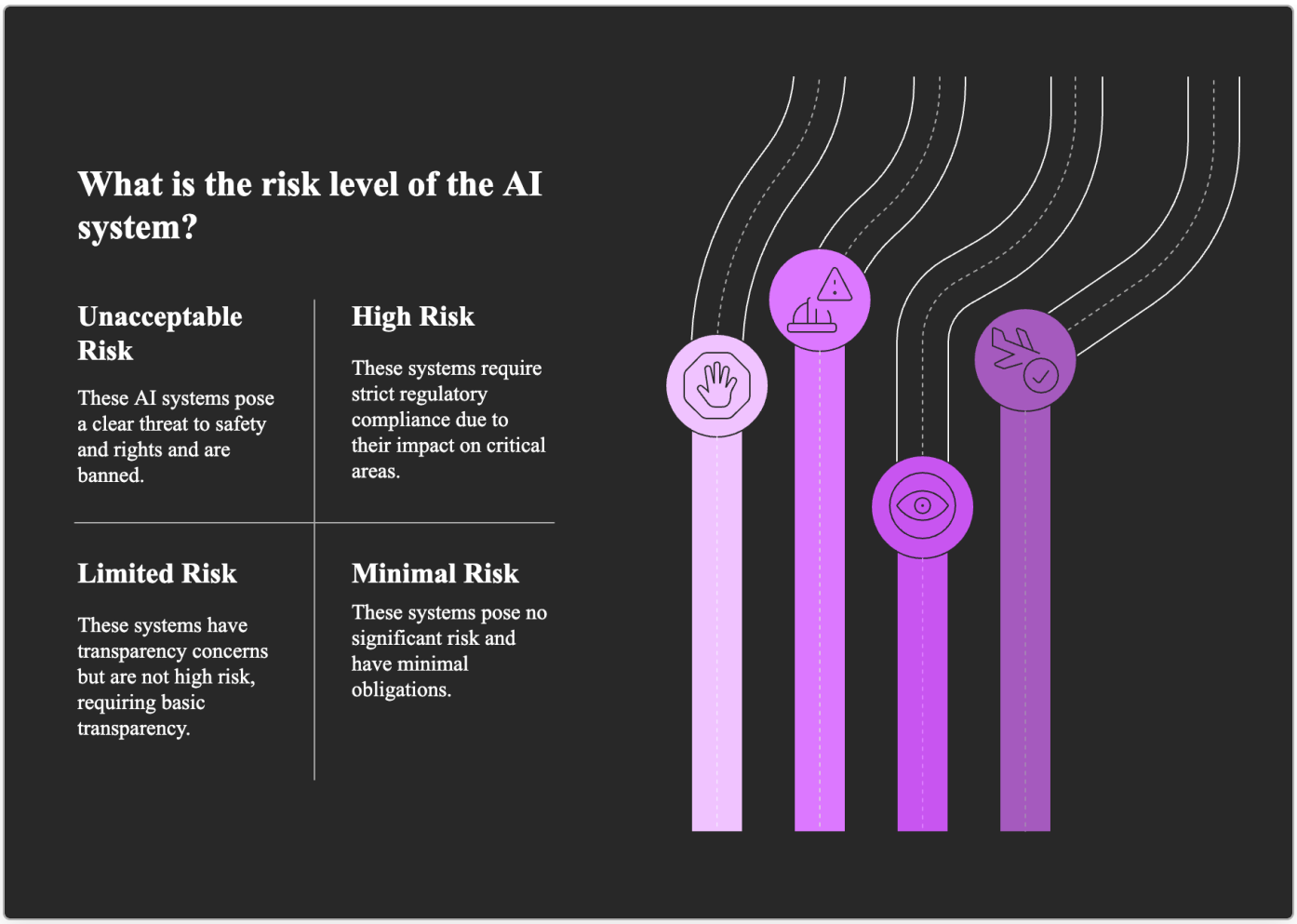


Figure: The EU AI Act employs a tiered risk classification framework to categorize AI systems into four levels – Unacceptable (red, e.g., social scoring), High (yellow, e.g., applications in healthcare or law enforcement), Limited (green, e.g., chatbots requiring transparency), and Minimal risk (blue, e.g., video game AI). Higher-risk categories face stricter compliance obligations and penalties for violations[5].

NIST AI Risk Management Framework (RMF)

In the United States, a prominent non-regulatory framework is the NIST AI RMF 1.0, published in January 2023. Developed through a multi-stakeholder process, the NIST AI RMF is a voluntary guidance for organizations to manage AI risks and promote trustworthy AI[6]. It provides a structured approach organized into four core functions: Govern, Map, Measure, and Manage.

- **Govern:** Establish organizational governance processes to oversee AI risk management (cultivating a culture of risk awareness, accountability, and adherence to

trustworthiness principles at all levels)[7]. Governance is a cross-cutting function that informs all others.

- **Map:** Contextualize and identify risks – i.e. understand the AI system's purpose, scope, and environment to recognize what risks might arise and who might be affected.
- **Measure:** Analyze, assess, and monitor AI risks – for example, measure the performance of mitigations, track metrics like bias or robustness, and audit the AI system to gauge if risk controls are effective[8].
- **Manage:** Mitigate and respond – implement controls to address identified risks (e.g. retraining a model on more diverse data to reduce bias, or enforcing human review for certain AI decisions), and have processes to respond to incidents or adapt the AI system as its context changes[8].

These functions operate in a continuous, iterative cycle (much like the cybersecurity framework's identify/protect/detect/respond/recover). The NIST framework is technology-neutral and use-case agnostic, meaning it can be applied to any AI system to improve its trustworthiness. It emphasizes stakeholder engagement, transparency, fairness, and other "qualities of trustworthy AI" as cross-cutting principles. While not law, the NIST AI RMF has been influential globally – for example, companies might use it as a basis for internal AI policies, and it aligns with ISO 23894's risk management approach[2] [1]. In practice, an organization using the NIST AI RMF would document the context of each AI application (Map), perform risk assessments and impact evaluations (Measure), apply safeguards and controls (Manage), and have an overarching governance program to tie these together (Govern). NIST has also released a companion AI RMF Playbook with actionable guidance and an AI RMF Crosswalk mapping its recommendations to other standards[6]. The goal is to help organizations "incorporate trustworthiness considerations into the design, development, use, and evaluation" of AI systems[6], thereby reducing risks and harms.

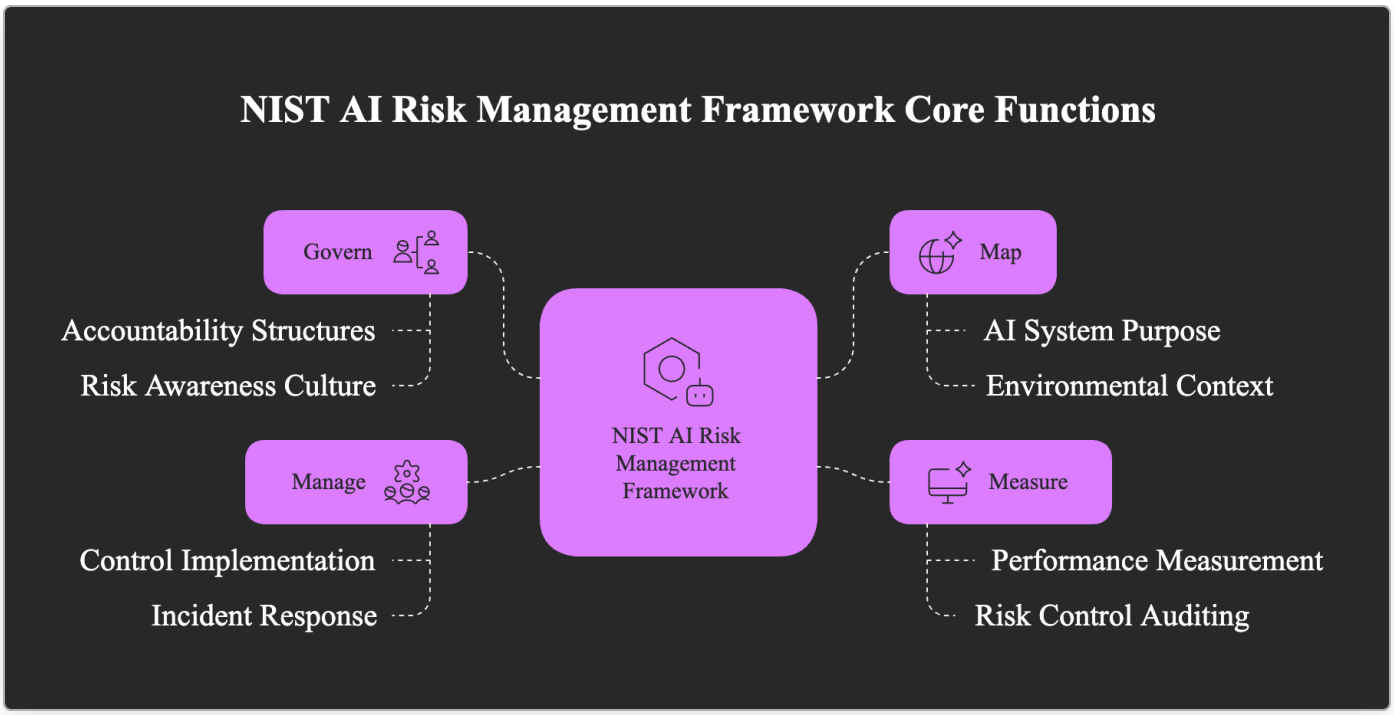


Figure: The NIST AI Risk Management Framework organizes AI risk management into four functions – Map (recognize context and identify risks), Measure (analyze and track risks), Manage (prioritize and mitigate risks), with Govern as an overarching function ensuring a risk culture and oversight[7]. This iterative process helps organizations build trustworthy AI by systematically addressing risks across the AI lifecycle.

Privacy and Data Protection Laws (GDPR, CCPA, etc.)

Data is the lifeblood of AI, so privacy laws are highly relevant to AI governance. The EU's General Data Protection Regulation (GDPR) imposes strict requirements on processing personal data, which many AI systems do (for instance, AI analyzing user behavior or personal attributes). GDPR mandates principles like data minimization, purpose limitation, and fairness in data processing[4]. It also provides individuals rights that affect AI – notably the right not to be subject to solely automated decisions with significant effects (Article 22 GDPR) in certain cases, or at least the right to human review and an explanation[9]. This means if a company uses an AI algorithm alone to decide something like a loan approval, an EU consumer may challenge it and request human intervention. AI governance programs must therefore incorporate GDPR compliance: ensuring lawful basis for AI data processing, conducting Data Protection Impact Assessments when deploying high-risk AI, and enabling transparency and recourse for individuals. GDPR's emphasis on accountability requires organizations to be able to demonstrate compliance – in context of AI, this entails

documenting how models were trained, how data was obtained and secured, and what measures prevent privacy violations[4]. Additionally, GDPR's fines for non-compliance (up to 4% of global turnover) make privacy a board-level risk.

In the US, while there is no federal GDPR equivalent, state laws like the California Consumer Privacy Act (CCPA) grant consumers rights over personal information. The CCPA (amended by CPRA) gives California residents the right to know, delete, and opt-out of the sale of their personal data[10], among other rights (like correct inaccurate data and not be discriminated against for exercising privacy rights)[10]. For AI, this means any models using California residents' data need processes to delete that data upon request and to stop selling or sharing it if the consumer opts out. CCPA also requires transparency in privacy policies about data usage which extends to AI-driven data uses. Further, sector-specific privacy rules may apply (e.g. HIPAA for health data used in AI medical diagnostics, which mandates strict safeguards for Protected Health Information).

Security and AI-Specific Regulations

Beyond privacy, AI governance intersects with cybersecurity and domain-specific regulations. AI systems must be secured against breaches and adversarial attacks (which is part of AI safety). Laws like the EU's Network and Information Security (NIS2) Directive or various national cybersecurity laws require organizations to protect systems (including AI) from cyber threats – failing to do so can cause data leaks or model manipulation. Moreover, industries have their own AI-related guidance. For example, the FDA in the US is developing regulatory guidance for AI/ML-based medical devices, requiring algorithm transparency and Good Machine Learning Practice for safety and effectiveness. In finance, regulators (like banking authorities) are examining algorithms for fairness and stability – the Federal Reserve's SR 11-7 guidance on model risk management, while older, is now being interpreted to include AI models (requiring rigorous validation and audit of models that affect financial decisions). The EU's draft AI Act explicitly covers AI in credit scoring, insurance, employment, etc. as high-risk, linking to existing financial regulations. Meanwhile, sectoral standards (like ISO 26262 for automotive functional safety) are relevant when AI is

used in that sector (e.g. self-driving car AI must meet both AI governance standards and functional safety standards).

Ethical Frameworks and Global Principles

Apart from hard law and formal standards, numerous ethical AI frameworks guide governance. The OECD AI Principles (2019) – adopted by 40+ countries – articulate high-level tenets: inclusive growth, human-centered values and fairness, transparency and explainability, robustness and safety, and accountability for AI[11] [12]. These principles have informed regulations and corporate codes of conduct worldwide. Similarly, the UNESCO Recommendation on AI Ethics (2021) provides a global ethical framework emphasizing human dignity, environmental well-being, gender equality, and peaceful use of AI. While these are not directly enforceable, they shape policymaker and public expectations, and organizations often voluntarily align with them to demonstrate commitment to responsible AI. For example, a company might commit to the OECD principle of robustness by establishing rigorous testing and validation before deploying AI, or to accountability by providing appeal mechanisms for AI decisions. In the U.S., the White House's Blueprint for an AI Bill of Rights (2022) lists principles echoing these themes (safe and effective systems, algorithmic discrimination protection, data privacy, notice & explanation, and human alternatives)[13]. Even though it's guidance, it sets a benchmark for what citizens should expect from AI – and thus what companies should strive to implement.

The regulatory environment for AI is multifaceted: companies must navigate international standards (ISO/IEC), follow region-specific laws (like the EU AI Act's risk-based rules and data protection laws such as GDPR/CCPA), and heed industry-specific requirements. Compliance will require cross-functional effort: lawyers and compliance officers to interpret laws, data scientists and engineers to implement technical controls, and executives to integrate these requirements into corporate governance. Many organizations are adopting AI governance frameworks that "ensure consistency and coherence" amidst this patchwork[1], often using standards like NIST and ISO as foundational, then layering on the specific legal obligations for their industry and markets.

Privacy, Data Governance, and Security Considerations

Privacy and data governance are critical pillars of AI governance because AI systems often consume and generate massive amounts of data, including personal and sensitive information. Ensuring compliance with privacy laws (as discussed with GDPR and CCPA) is just the starting point. Organizations must also institute strong data governance practices to maintain data quality, protect confidentiality, and prevent bias. This includes:

Data Quality & Lineage

AI outcomes are only as good as the data they are trained on. Good data governance means tracking the provenance of data, assessing its quality, and ensuring representativeness. Poor data (e.g. skewed demographics in training data) can lead to biased AI decisions. Organizations should maintain data documentation, schemas, and audit trails to know what data went into each model. Some are adopting datasheets for datasets (a practice proposed by Gebru et al.)[19] to catalogue the characteristics and potential limitations of data used in AI. This helps in risk assessment and in explaining AI decisions (as required by transparency mandates).

Data Minimization & Access Control

In line with privacy principles, AI systems should use the minimum data necessary for their purpose. Personal data that is not needed should not be collected; if it is needed, techniques like pseudonymization or encryption should secure it. Role-based access controls and data usage policies are necessary so that only authorized personnel or processes can access sensitive data (reducing insider risk or accidental misuse). For AI, an extra layer involves controlling access to model inputs and outputs if they could reveal sensitive information (for instance, large language models can inadvertently memorize and output personal data seen in training, so governance should restrict use of training sets with personal data or employ techniques to sanitize it).

Privacy-Enhancing Technologies

Technical measures such as differential privacy, federated learning, and homomorphic encryption are increasingly part of AI governance. These allow AI models to learn from data without exposing individual data points. For example, a machine learning model can be trained across distributed datasets held by different hospitals using federated learning, so that sensitive health data never leaves the hospital premises, complying with privacy regulations. Differential privacy can add statistical noise to AI model outputs to protect any single individual's data from being reverse-engineered. While not mandated by law, these techniques demonstrate a commitment to privacy and can enable the use of data that would otherwise be off-limits due to regulation.

Retention and Purpose Limitation

Data governance policies should define how long data used in AI is retained and for what purposes it can be reused. GDPR, for instance, requires that data not be kept longer than necessary. For AI, retaining historical training data or model outputs without limit can become a liability (both in terms of storage risk and regulatory risk). Many organizations set retention schedules and deletion procedures, and also ensure that if an AI model is repurposed, it doesn't inadvertently use data in ways that go beyond the original consent or purpose (addressing the purpose limitation principle).

On the security front, AI systems introduce new dimensions to cybersecurity and IT risk management:

Model Security

AI models themselves can be targets of attack. Adversaries might try to steal a model (to copy a company's valuable IP) or exploit it via adversarial examples (feeding inputs that cause the model to err). Robust AI governance includes securing the model binaries and APIs, using adversarial training or input filtering to harden models against manipulation, and monitoring for model drift or anomaly inputs. For example, an image recognition AI might be vulnerable to specially crafted images; governance would call for testing the model against such inputs and possibly implementing runtime defenses.

Data Security

The data pipelines feeding AI must be secured. This overlaps with general cybersecurity: encrypt data in transit and at rest, use strong identity management for systems accessing training data, and apply intrusion detection on systems where AI data is stored. The concern is both confidentiality (preventing breaches of sensitive data) and integrity – if an attacker can poison your training data (tamper with it), they could subtly corrupt the AI's behavior. Thus, AI governance should encompass data validation and integrity checks. Some organizations version-control their training datasets and use hash checks to detect alterations.

Third-Party and Supply Chain Risks

Many AI solutions rely on third-party components – pre-trained models, cloud AI services, open-source libraries. These introduce supply chain risk. Governance needs to evaluate the provenance and security of third-party AI tools (e.g. ensure an open-source model doesn't have hidden backdoors, or that cloud providers have robust security certifications). Contractual agreements with AI vendors should include clauses on data security, incident notification, and compliance with relevant laws.

Incident Response

Despite preventive measures, things can go wrong – an AI system might cause an unforeseen incident (e.g. a self-driving car accident, or an AI chatbot that leaks confidential info). Organizations should extend their incident response planning to AI-specific scenarios. This means having protocols to quickly shut down or patch AI systems that behave unexpectedly, processes to communicate with stakeholders and regulators if an AI causes harm or a data breach occurs through an AI component, and forensics capabilities to investigate AI-related incidents. Some regulators (like the EU AI Act) will likely mandate reporting certain AI "malfunctions" or incidents. Even when not mandated, it's a good practice to treat AI incidents with the seriousness of safety or security incidents.

Ethical Hacking and Auditing

Just as penetration testing is common for cybersecurity, AI governance can include "red-teaming" AI models – having experts attempt to trick or defeat the model's safeguards to identify weaknesses. For instance, before deploying a content moderation AI, an organization might hire a team to find inputs that evade the filter (to improve it). Regular audits of AI systems against criteria like fairness, privacy, and security are becoming best practice (and may be required under certain laws or certifications).

Finally, privacy, security, and data governance concerns are often intertwined. For example, data breaches can lead to both security incidents and privacy violations. A robust AI governance program will therefore coordinate across these domains – ensuring that the Chief Information Security Officer (CISO) or security team is involved in AI projects, that the Chief Privacy Officer and data protection officers have oversight of AI data usage, and that data governance committees include representation from AI development teams. Many companies find it useful to establish an AI/ML governance council that brings together legal, compliance, privacy, security, and AI technical leads to create unified policies. The outcome is that AI systems are not treated in isolation but are integrated into the organization's overall risk management for data and IT.

Technical Aspects of AI Safety and Governance

AI governance is not just about high-level policies; it also involves concrete technical measures and best practices to ensure AI systems are safe, reliable, and aligned with intended goals. We highlight key technical aspects and tools that AI practitioners and risk managers employ as part of governance:

Robustness and Reliability

Technical AI safety begins with making models robust to errors and uncertainties. This involves rigorous validation of AI models on test data that simulates real-world variability and potential edge cases. Techniques like stress testing are used – e.g. testing a computer vision system in varying lighting conditions or with noise – to ensure the model still performs acceptably. Robustness research also addresses adversarial examples[14], which are inputs intentionally designed to fool AI (like subtly altering a stop sign image so an AI misreads it).

To govern against this, organizations might incorporate adversarial training (training the model on adversarial samples so it learns to resist them) and incorporate redundancies (multiple sensors or ensemble models) so that one component catching an error can compensate for another. The goal is to avoid AI failures in deployment, especially for safety-critical systems (like AI in cars, medical diagnosis, or critical infrastructure). Governance frameworks often require documenting the operating domain of an AI and its known failure modes, and ensuring it's only used under conditions it was designed for. For example, an AI model valid for English text shouldn't be applied to French text without retraining – a governance check would prevent such misuse.

Alignment with Human Values

Alignment is a technical AI safety field focused on ensuring that AI's objectives and behavior remain in line with human intentions and values[14]. For current AI systems, this can mean incorporating human feedback and ethical considerations during development. One approach is Reinforcement Learning from Human Feedback (RLHF), famously used to align large language models (like ChatGPT) with desirable behavior by training them on examples of good and bad responses. Alignment also involves specifying constraints – for instance, an AI scheduling tool might be aligned with fairness values by adding rules that it not systematically give one group the worst time slots. As AI systems become more autonomous or complex, alignment techniques become even more important to avoid "goal misalignment" where an AI optimizes something harmful because it misunderstands the human's true intent. Even simple ML models benefit from alignment thinking: define the objective function carefully (not just optimizing profit, but profit subject to fairness constraints, for example). Technical governance includes peer review of model objectives and metrics to catch misalignment early. In research contexts, alignment also refers to preparing for future advanced AI (AGI) to ensure it remains beneficial – while that may be beyond the scope of most organizations today, it underlines the principle that AI should remain under meaningful human control.

Interpretability and Transparency Tools

To build trust and enable oversight, technical teams use interpretability techniques that shed light on "black box" AI models[14]. This includes explainable AI (XAI) methods like SHAP values or LIME that indicate which features of an input most influenced a model's decision[21]. For neural networks, visualization tools might highlight which parts of an image a convolutional network focused on. Interpretability is important for debugging models (ensuring they're making decisions for the right reasons) and for explaining outcomes to stakeholders (like providing a reason for a loan denial). Some regulations (e.g. EU AI Act, GDPR's notion of explanation) implicitly push for such capabilities. Governance programs often require that for high-impact AI, an explanation report or model card be produced. Model Cards are a documentation technique describing a model's intended use, performance, and limitations in plain language – a practice recommended by Google and others to increase transparency[18]. By using these tools, organizations can "understand what's going on" inside AI[14], which aids accountability (engineers can justify that the model is making decisions based on legitimate factors rather than prohibited ones like race or gender, unless legally allowed and appropriate).

Bias and Fairness Mitigation

A major aspect of AI safety (in the societal sense) is ensuring AI does not unfairly discriminate or produce biased outcomes against protected groups. Technically, this involves measuring bias – e.g. checking model error rates or decision distributions across demographics – and then mitigating it. Techniques include pre-processing (ensuring training data is balanced or reweighted), in-processing (using algorithms that constrain the model to treat groups fairly, such as by adding fairness penalty terms in the objective), and post-processing (adjusting model outputs to reduce disparity). For instance, an HR resume screening AI might be audited for gender bias; if found, the team might retrain it without certain problematic features or use a fairness-aware learning method. Many organizations now use bias audit toolkits (some open-source, like IBM's AI Fairness 360 or Microsoft's Fairlearn) as part of model development and validation. AI governance can formalize this by mandating a "fairness check" before deployment of certain AI models, with documented results. In some jurisdictions, this is becoming law – e.g. New York City's Local Law 144 requires bias audits for AI hiring tools. Thus, technical and compliance aspects meet:

engineers must produce evidence (metrics, plots) that the model meets fairness thresholds, and compliance officers ensure those results pass legal muster.

Performance Monitoring and Drift Detection

Launching an AI model is not a one-and-done event. Governance requires ongoing monitoring to ensure the model remains safe and effective over time. Model drift occurs when data patterns change (for example, consumer behavior shifts or new types of inputs appear) such that the model's performance degrades. Technical measures involve tracking key performance indicators in production – e.g. accuracy, false positive/negative rates, or business metrics impacted by the AI – and setting up alerts if they move beyond acceptable ranges. If an image classifier that used to have 95% accuracy suddenly drops to 85%, that could indicate drift or an issue needing retraining. Additionally, monitoring for out-of-distribution inputs – data that is unlike what the model saw in training – is important. Techniques like anomaly detection can flag when the AI is asked to make predictions on data that may be outside its expertise, so that it can hand off to a human or at least signal low confidence. Good governance establishes a feedback loop: when the monitoring signals a problem, the organization has a process to respond (perhaps captured in the "Manage" function of the NIST RMF). This could involve retraining the model with fresh data, tuning it, or even rolling back to a previous model version.

Safety Constraints and Testing in Simulation

For AI that interacts with the physical world (robots, autonomous vehicles, medical devices), safety testing is crucial. This often means extensive simulation testing to explore scenarios that are rare or dangerous in the real world. Autonomous vehicle AI, for example, is tested in simulated environments for billions of miles to see how it reacts to every conceivable situation (child running into road, unusual traffic pattern, sensor failures, etc.). Governance would require meeting certain safety targets (e.g. proving the AI drives more safely than an average human) before expanding deployments. For robots, formal verification methods might be applied to certain decision-making modules to mathematically prove that, say, the robot will not exceed certain force limits around humans. Additionally, redundancy and fail-safes are technical safety measures: if the AI fails, there is a backup system to take over or

safely shut it down (in aviation, think of autopilot vs. pilot control). Many of these are well-known in traditional engineering, and now they are being applied to AI control systems.

In cutting-edge AI, technical safety research also looks at verification and validation of complex models (how to assure a neural network does what it should and nothing more), and controllability (ensuring we can intervene or shut down AI if needed, often called "safe fail" or "graceful degradation"). Some AI systems include an internal "guardian" algorithm that monitors the main AI and can override decisions if they seem unsafe, akin to a supervisory control.

By implementing these technical measures – robustness, alignment, interpretability, bias mitigation, monitoring, and rigorous testing – organizations build the technical foundations of trustworthy AI. However, these techniques must be supported by the governance processes discussed earlier (policies, roles, audits) to be effective. For example, there's little point in generating an interpretability report if no governance process requires reviewing it for potential problems. Thus, technical and organizational aspects of AI governance work hand in hand. An AI practitioner might focus on these techniques in their daily work, while a compliance officer or executive ensures that the work is done and acted upon as part of a broader risk management strategy.

Audience-Specific Guidance: Focus and Concerns

Different stakeholders in an organization have distinct roles in AI governance. Here we provide tailored guidance for AI Practitioners, Compliance Officers, Executives, and Policymakers, addressing what each group should focus on and their key concerns:

AI Practitioners

Compliance Officers

Data Scientists, ML Engineers, AI Developers

Focus:

AI practitioners are at the front lines of building and deploying AI systems. Their focus is on integrating governance and safety principles into the technical development process. This means operationalizing Responsible AI on a day-to-day basis. Practitioners should aim to build models that not only perform well on accuracy metrics, but also meet criteria for fairness, explainability, and robustness.

Specific Concerns and Practices:

- **Implementing Ethical Guidelines:** Practitioners should be familiar with their organization's AI ethics or responsible AI guidelines and translate them into concrete actions. For example, if the guideline is to ensure fairness, the practitioner needs to choose appropriate algorithms or add bias mitigation steps (as discussed in the technical

Legal, Regulatory, and Ethics Compliance Personnel

Focus:

Compliance officers are concerned with ensuring that AI systems and processes adhere to all relevant laws, regulations, and internal policies. Their focus is on oversight, auditing, and guiding the organization's AI efforts from a risk and compliance perspective. They translate regulatory requirements (like those in GDPR, AI Act, etc.) into controls and checklists that the technical teams should follow, and they verify those controls are met.

Specific Concerns and Responsibilities:

- **Regulatory Monitoring:** Compliance officers need to stay on top of the fast-changing AI regulatory environment. They should track laws like the EU AI Act, FTC guidelines (in the US context), industry-specific rules (e.g. health AI guidelines from FDA or European Medicines Agency), and even soft law (like

section) during model development. If transparency is a principle, they should incorporate explainability tools and be prepared to produce documentation like model cards. Essentially, they "translate principles into practice"[15].

- **Data Handling and Privacy:** Practitioners often collect and preprocess data – they must ensure compliance with data governance rules. This includes anonymizing data when required, obtaining proper consent or legal basis for data use, and respecting opt-outs (e.g. filtering out data of users who withdrew consent). If using personal data, involving privacy experts and conducting privacy impact assessments is key. Also, practitioners should use secure data storage and coding practices to avoid leaks (security is part of their responsibility too).
- **Testing and Validation:** A core concern is releasing a model that later behaves unpredictably. Practitioners should therefore invest effort in exhaustive testing – not just standard train/test

ethics certifications or standards such as ISO 42001). Understanding these frameworks allows them to interpret what is required for their organization. For example, if the EU AI Act classifies an AI product as high-risk, the compliance officer must ensure a conformity assessment (possibly involving a notified body) is planned and that all documentation (technical file) is ready to demonstrate compliance (risk assessment, data governance, transparency, accuracy testing, etc., as required by the Act).

- **Policy Development:** Compliance roles often include developing internal policies or SOPs for AI. This could mean writing an AI Governance Policy that states how AI projects are evaluated and approved, what ethical principles must be followed, and how to handle incidents. They might also help draft AI usage guidelines for third-party AI tools, ensuring due diligence is done (for instance, prohibiting use of an external AI

splits, but stress tests, adversarial tests, and corner-case analysis. Governance might require peer review of models; practitioners should be ready to have their models audited or reproduced by colleagues (ensuring code and data are well-managed for reproducibility). Maintaining an AI model inventory is useful – practitioners document each model's purpose, training dataset, version, and outcomes of validation checks. This inventory, often mandated by governance, helps track compliance and performance over time.

- **Continuous Monitoring:** After deployment, practitioners might be responsible for monitoring AI in production (especially in smaller organizations without a separate ML-Ops team). They should set up dashboards or alerts for model performance and be proactive in retraining or tuning models when drift is detected. Their concern is to prevent "silent failures" where an AI's accuracy degrades unnoticed, possibly causing harm or errors (like a loan approval service that hasn't been vetted for privacy compliance).
- **Training and Awareness:** Compliance teams often organize training for developers, product managers, and other stakeholders on relevant compliance topics (privacy, anti-discrimination laws, etc.). For AI, they might introduce mandatory training on "AI Ethics & Compliance," covering issues like how to avoid discriminatory outcomes, how to do documentation, or what the law says about explainability. They need to cultivate awareness so that others in the organization recognize potential compliance issues early (e.g., a product manager realizing "we can't launch this feature without a user consent mechanism because it involves personal data profiling").
- **Review and Audit:** A core task is to review AI systems for compliance before and after deployment. This can involve checking that a Data Protection Impact Assessment (DPIA) is completed, ensuring contracts cover AI responsibilities, auditing

model slowly becoming biased as demographics shift).

- **Collaboration with Compliance:** Practitioners should view compliance and risk teams as partners, not adversaries. By engaging early with compliance officers or legal advisors, they can get clarity on constraints (e.g. "this model falls under high-risk AI per EU AI Act, so we need to implement X, Y, Z documentation and get a conformity assessment"). This prevents rework and ensures the product will be launchable in target markets. For instance, if a developer knows upfront that an AI tool will need to explain its decisions to users to meet a regulation, they can build in that functionality from the start.
- **Mindset and Culture:** Practitioners benefit from cultivating a "safety culture" akin to DevSecOps where security is everyone's job – here, ethical AI is everyone's job. They should feel ownership of not just delivering a working model, but delivering a responsible model. This might

algorithms for bias, and maintaining documentation.

- **Incident and Issue Handling:** If an AI system causes a potential compliance issue, compliance officers coordinate the response. They might need to report incidents to authorities, investigate violations, and ensure new controls are implemented.

Key Concerns:

Compliance officers are ultimately concerned about legal liability, regulatory sanctions, and reputational risk. They aim to prevent non-compliance fines by being proactive and enforcing thorough risk assessments. Bridging the gap between compliance requirements and technical implementation is a key challenge.

Compliance officers serve as the guardians of ethical and lawful AI use in the organization. They establish guardrails and perform checks, ensuring the company's AI initiatives do not run afoul of regulations and ethical norms.

involve speaking up if they notice an AI use case that could be harmful or if data is biased. Organizations can support this with training (e.g. on fairness in AI) and by not solely incentivizing speed and accuracy, but also quality in ethical terms.

AI practitioners ensure that the code and models they produce have governance considerations “baked in,” addressing trust, safety, and compliance requirements from the ground up.

Executives & Board Members

C-Suite Executives and Board Directors

Focus:

Executives are responsible for the strategic oversight and organizational commitment to AI governance and responsible AI. They

Policymakers & Regulators

Government Officials and Regulatory Authorities

Focus:

Policymakers are responsible for creating and enforcing the rules that govern AI in society. Their focus is on addressing public risks and harms from AI, ensuring innovation benefits

focus on balancing innovation with risk management and maintaining the company's reputation and trust with stakeholders. Executives need to ensure that appropriate resources, culture, and structures are in place for effective AI governance. In many cases, this means setting the tone at the top that AI ethics and compliance are priorities, not optional.

Specific Concerns and Roles:

- **Strategy and Investment:** Leaders integrate AI governance into the overall business strategy, allocate resources for mitigation, and decide on risk appetite.
- **Governance Structures:** Executives create and empower governance bodies like AI Governance Boards or steering committees and formalize roles related to AI oversight.
- **Culture and Tone:** Executives set the tone that ethical AI is core to the company's mission, communicating its importance and rewarding responsible behavior.
- **Accountability and Risk Oversight:** Executives are

society while protecting rights and safety. They must balance encouraging technological progress with mitigating potential negative impacts.

Specific Focus Areas:

- **Developing AI Regulations and Standards:** Policymakers work on frameworks like the EU AI Act and sector-specific rules, identifying high-risk uses and appropriate safeguards.
- **Harmonization and International Cooperation:** They work in international forums (OECD, G20, UNESCO) and standard-setting organizations (ISO, IEEE) to harmonize approaches globally.
- **Enforcement Mechanisms:** Regulators focus on developing methods to audit algorithms, investigate complaints, sanction non-compliance, and possibly create new oversight bodies (e.g., EU AI Board).
- **Addressing Societal Concerns:** They consider issues like job displacement, AI literacy, and might fund responsible AI

accountable for AI outcomes, monitor governance performance through metrics and reports, and manage AI-related crises.

- **Compliance with Emerging Regulations:** Executives ensure the organization is prepared for upcoming laws like the EU AI Act and may seek certifications (like ISO 42001).
- **Opportunity and Innovation:** Leaders see responsible AI as a market differentiator and incorporate “Trustworthy AI” into branding, potentially sponsoring innovation in governance tools.

Executives also consider broader societal impacts and alignment with CSR/ESG commitments.

Executives need to champion AI governance as an integral part of doing business in the AI era. They ensure the alignment of people, processes, and technology for responsible AI, viewing it as essential for long-term success and trust.

innovation or set high standards for government AI use.

Concerns:

Primary concerns include preventing harm, preserving fundamental rights (privacy, non-discrimination), ensuring national security, and maintaining AI transparency and accountability[16]. They also balance these with fostering innovation, avoiding over-regulation that might stifle progress.

Policymakers and regulators create the external requirements and incentives for AI governance, addressing macro-level issues to protect public interest and create a level playing field.

AI Maturity Stages Frameworks (Seven-Stage Progression)

To help organizations assess and improve their AI governance and responsible AI practices, we present six AI Maturity Stages frameworks. Each is a seven-stage model describing the progression from rudimentary or non-existent capabilities to highly advanced and integrated capabilities in a specific area of AI governance. The areas are: AI Governance, AI Safety, AI Trust & Transparency, Responsible AI, AI Risk Management, and AI Compliance. Organizations can evaluate which stage best describes them in each area and identify steps to advance to higher maturity. Each stage includes key characteristics, assessment criteria, best practices to implement, and typical challenges to overcome. While the details differ for each area, a common theme is moving from an ad-hoc, reactive approach toward a proactive, optimized, and continuously improving approach.

Below, we outline each maturity Stages and its seven stages:

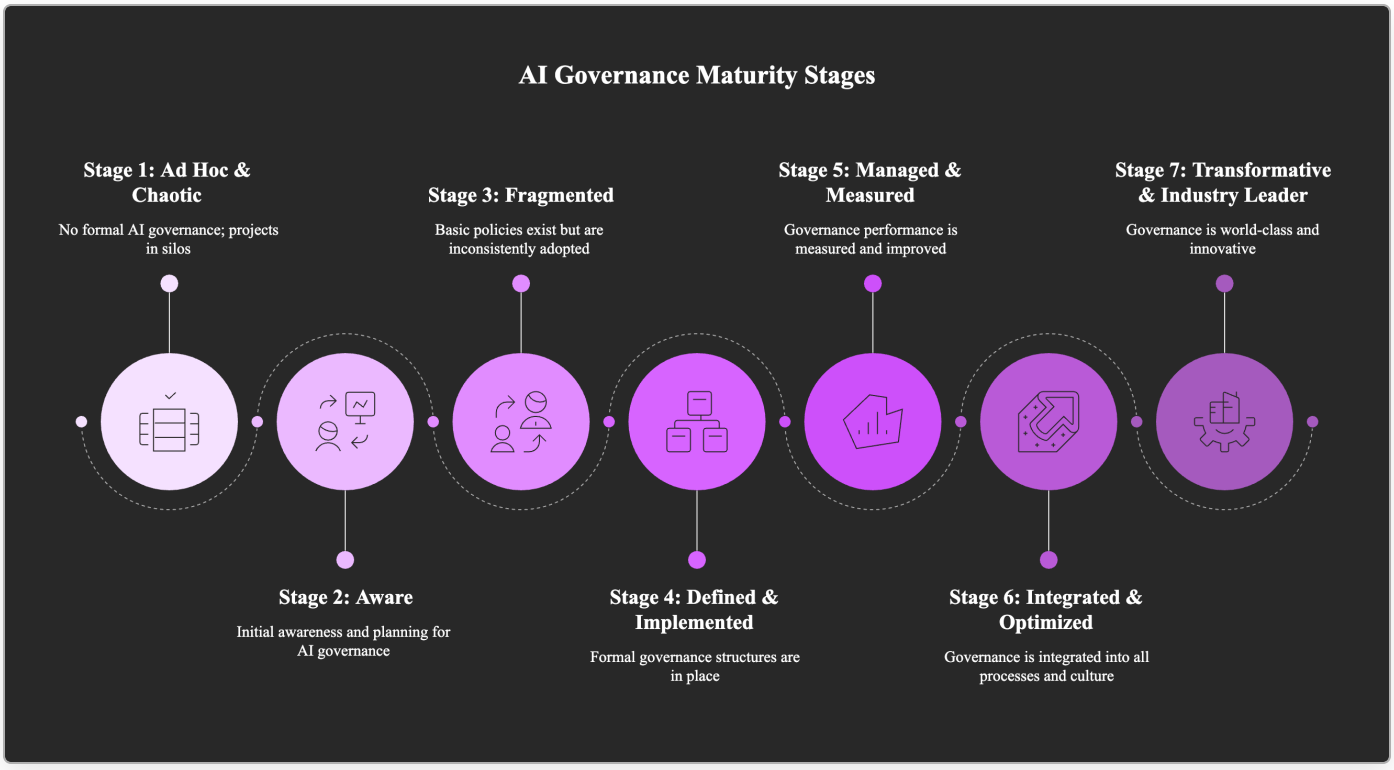


Figure: 7 States of AI Governance Maturity.

1. AI Governance Maturity Stages (Organizational AI Governance Capability)

Stage 1: Ad Hoc & Chaotic ▶

Stage 2: Aware (Initial Awareness & Planning) ▶

Stage 3: Fragmented (Basic Policies, Inconsistent Adoption) ▶

Stage 4: Defined & Implemented ▶

Stage 5: Managed & Measured ▶

Stage 6: Integrated & Optimized ▶

Stage 7: Transformative & Industry Leader ▶

Organizations can use this AI Governance maturity model to pinpoint where they stand (e.g. maybe Stage 3 if they have some policies but inconsistent application) and plan targeted improvements to progress (e.g. to Stage 4 by formalizing a governance committee and mandating processes across all projects).

2. AI Safety Maturity Stages (Technical Safety & Reliability of AI Systems)

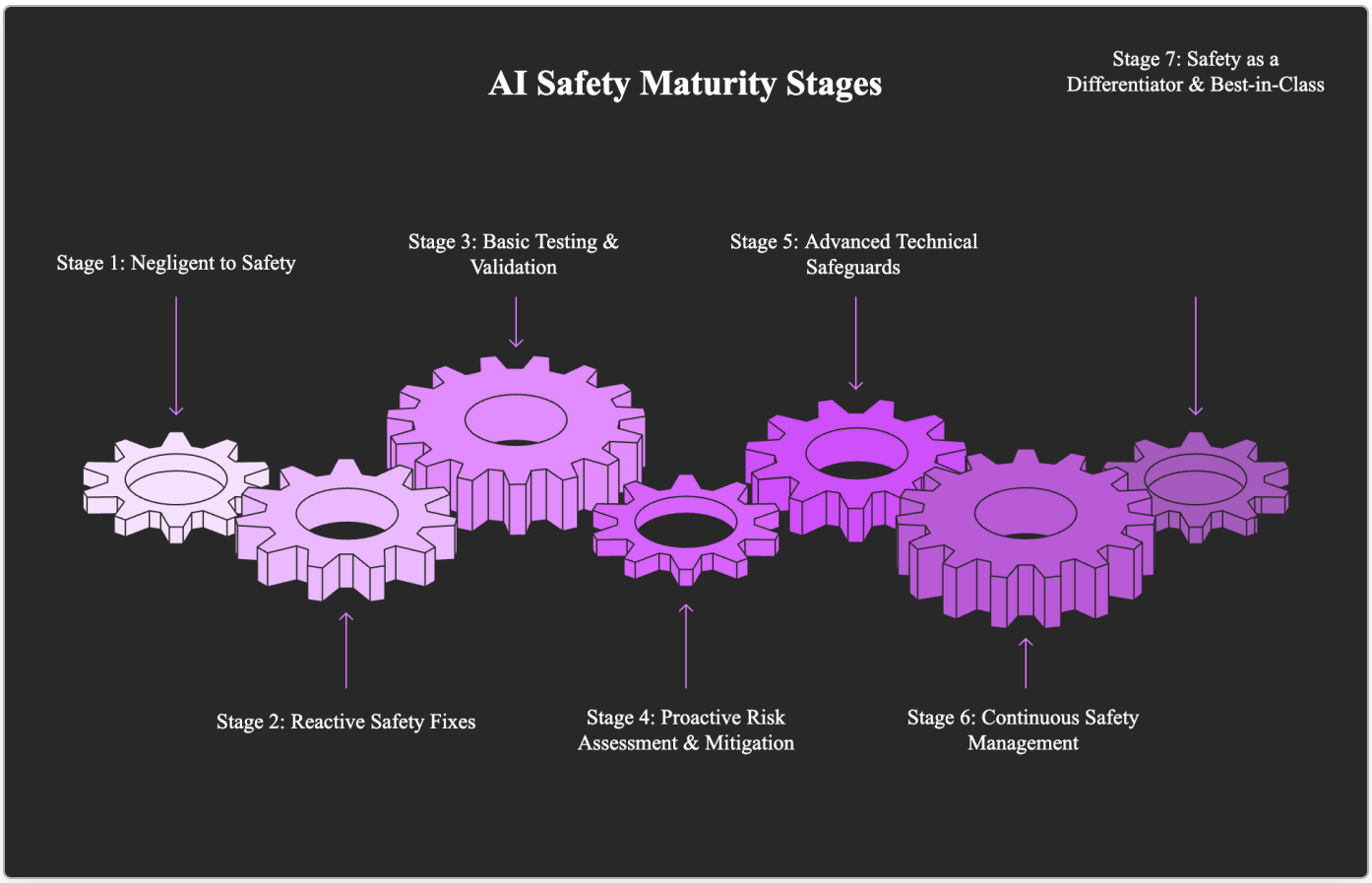


Figure: AI Safety Maturity Stages

Stage 1: Negligent to Safety ▶

Stage 2: Reactive Safety Fixes ▶

Stage 3: Basic Testing & Validation ▶

Stage 4: Proactive Risk Assessment & Mitigation ▶

Stage 5: Advanced Technical Safeguards ▶

Stage 6: Continuous Safety Management ▶

Stage 7: Safety as a Differentiator & Best-in-Class ▶

Using this maturity model, an organization can evaluate how well it currently handles AI safety. For example, a fintech startup might realize they're at Stage 2 (only reacting when models behave badly) and set a goal to reach Stage 4 by implementing proactive risk assessments and more rigorous testing in the next year. Each step up in maturity significantly reduces the likelihood of catastrophic failures and builds trust in the AI systems both internally and with clients/regulators.

3. AI Trust & Transparency Maturity Stages (Building Stakeholder Trust and Providing Transparency)

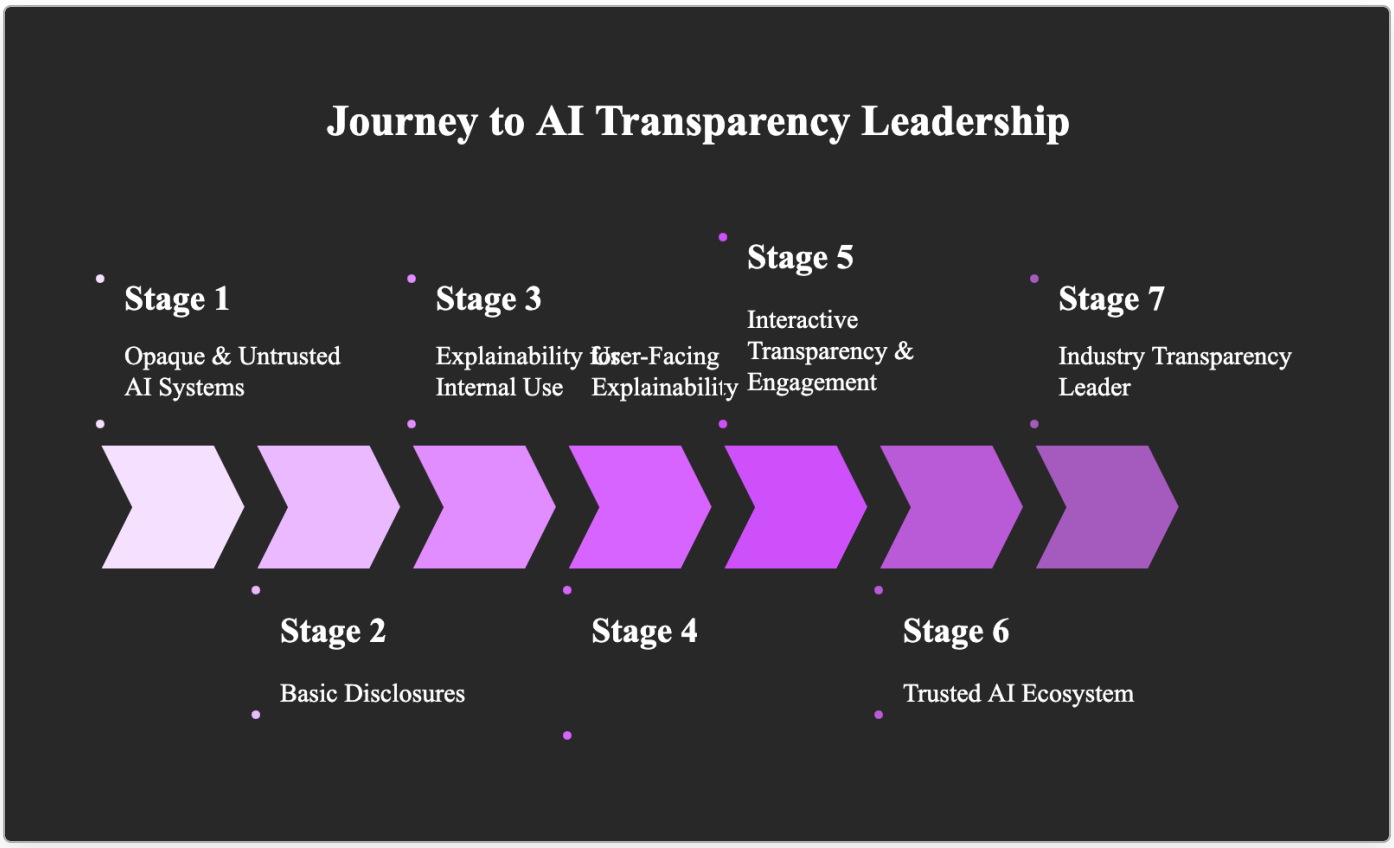


Figure:Trust & Transparency Stages.

Stage 1: Opaque & Untrusted ▼

AI systems are essentially black boxes, with no efforts at transparency or explanation. Users and other stakeholders are kept in the dark about when AI is used or how decisions are made. As a result, there is suspicion or low trust; if something goes wrong, the default assumption might be the AI is at fault.

Assessment: No documentation provided to users, no model cards, no explainability tools used. Possibly you find users complaining "I got this result and I have no idea why."

Challenge: Low trust can lead to user pushback or non-adoption. Also regulators might intervene if they find lack of transparency problematic (especially in regulated sectors).

Best Practice: Start with disclosure – at minimum, tell people when they are interacting with or subject to an AI decision (e.g. a simple statement: "This decision was generated by an algorithm."). This aligns with emerging norms like the AI Act's transparency requirements for chatbots or deepfakes.

Stage 2: Basic Disclosures ▶

Stage 3: Explainability for Internal Use ▶

Stage 4: User-Facing Explainability ▶

Stage 5: Interactive Transparency & Engagement ▶

Stage 6: Trusted AI Ecosystem ▶

Stage 7: Industry Transparency Leader ▶

This maturity Stages helps in assessing how well an organization fosters trust through transparency and engagement. For example, a social media company deploying AI for content filtering might realize it's at Stage 2 (just basic disclaimers) but facing public distrust. They could aim for Stage 4 by implementing user-visible explanations for why a post was taken down and providing a way to appeal or get more info. As they climb stages, they should see trust metrics improve – which can correlate with user satisfaction and loyalty.

4. Responsible AI Maturity Stages (Ethical and Social Responsibility in AI Use)

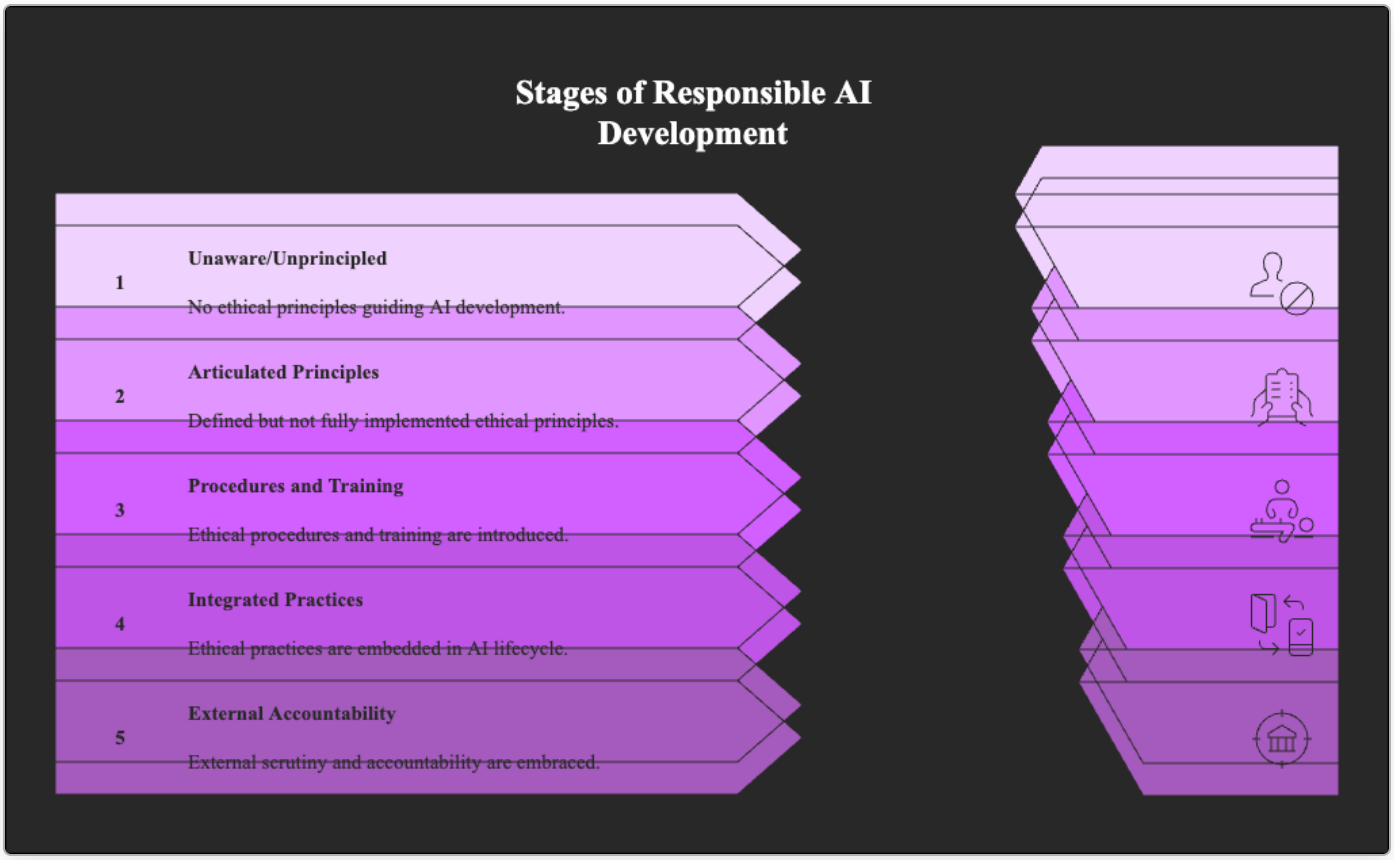


Figure: Stages of Responsible AI Maturity.

Stage 1: Unaware/Unprincipled ▶

Stage 2: Articulated Principles (on Paper) ▶

Stage 3: Procedures and Training for Ethics ▶

Stage 4: Integrated Responsible AI Practices ▶

Stage 5: External Accountability and Audit ▶

Stage 6: Culture of Responsibility & Empowerment ▶

Stage 7: Social Stewardship and Advocacy ▶

This maturity model is inspired by general ethical governance maturity models (similar to those used in corporate social responsibility and the GSMA's Responsible AI roadmap[17]). It helps an organization gauge its commitment to not just doing AI right, but doing the right AI. For example, a company at Stage 2 (principles on website) might have faced criticism that those aren't in practice. To move to Stage 3 and 4, they'd start implementing oversight and integrating ethics into workflows. The higher stages (5-7) become relevant for large organizations that wish to maintain public trust at scale and influence the wider ecosystem.

5. AI Risk Management Maturity Stages (Holistic Management of AI Risks)

AI Risk Management Maturity Stages

No AI-specific Risk Management

AI risks are not distinguished from general project risks.

Qualitative Acknowledgment of AI Risks

Major AI risks are qualitatively identified in key projects.

Structured Risk Assessment Process

A structured process for assessing AI risks is established.

Risk Mitigation and Control Implementation

Controls for identified AI risks are systematically implemented.

Integrated Risk Management & Monitoring

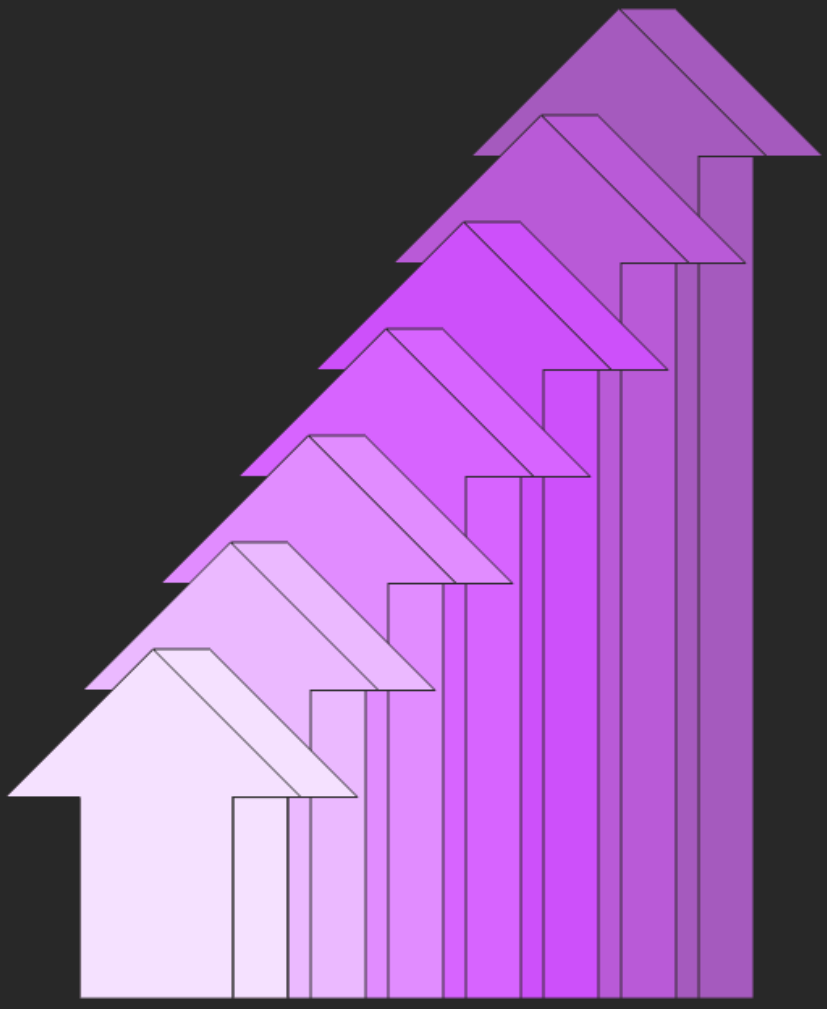
AI risk management is integrated into enterprise risk management.

Advanced Quantitative Risk Analysis

Advanced quantitative methods for AI risk analysis are employed.

Adaptive and Resilient Risk Posture

The organization develops a highly adaptive and resilient risk posture.



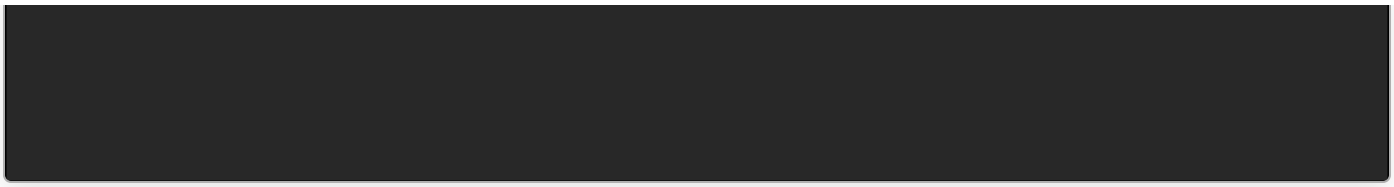


Figure: Stages of AI Risk Management Maturity.

Stage 1: No AI-specific Risk Management ▶

Stage 2: Qualitative Acknowledgment of AI Risks ▶

Stage 3: Structured Risk Assessment Process ▶

Stage 4: Risk Mitigation and Control Implementation ▶

Stage 5: Integrated Risk Management & Monitoring ▶

Stage 6: Advanced Quantitative Risk Analysis ▶

Stage 7: Adaptive and Resilient Risk Posture ▶

Using this model, an organization can measure maturity by seeing how formalized and effective their AI risk processes are. A bank might find they are at Stage 4 (they have assessments and controls, integrated in model risk management) but want to reach Stage 6 with more quantitative rigor due to regulatory expectations. A startup might be Stage 2 and need to move to 3 and 4 quickly as they scale to enterprise clients who demand evidence of risk controls.

6. AI Compliance Maturity Stages (Adherence to External Regulations & Standards)

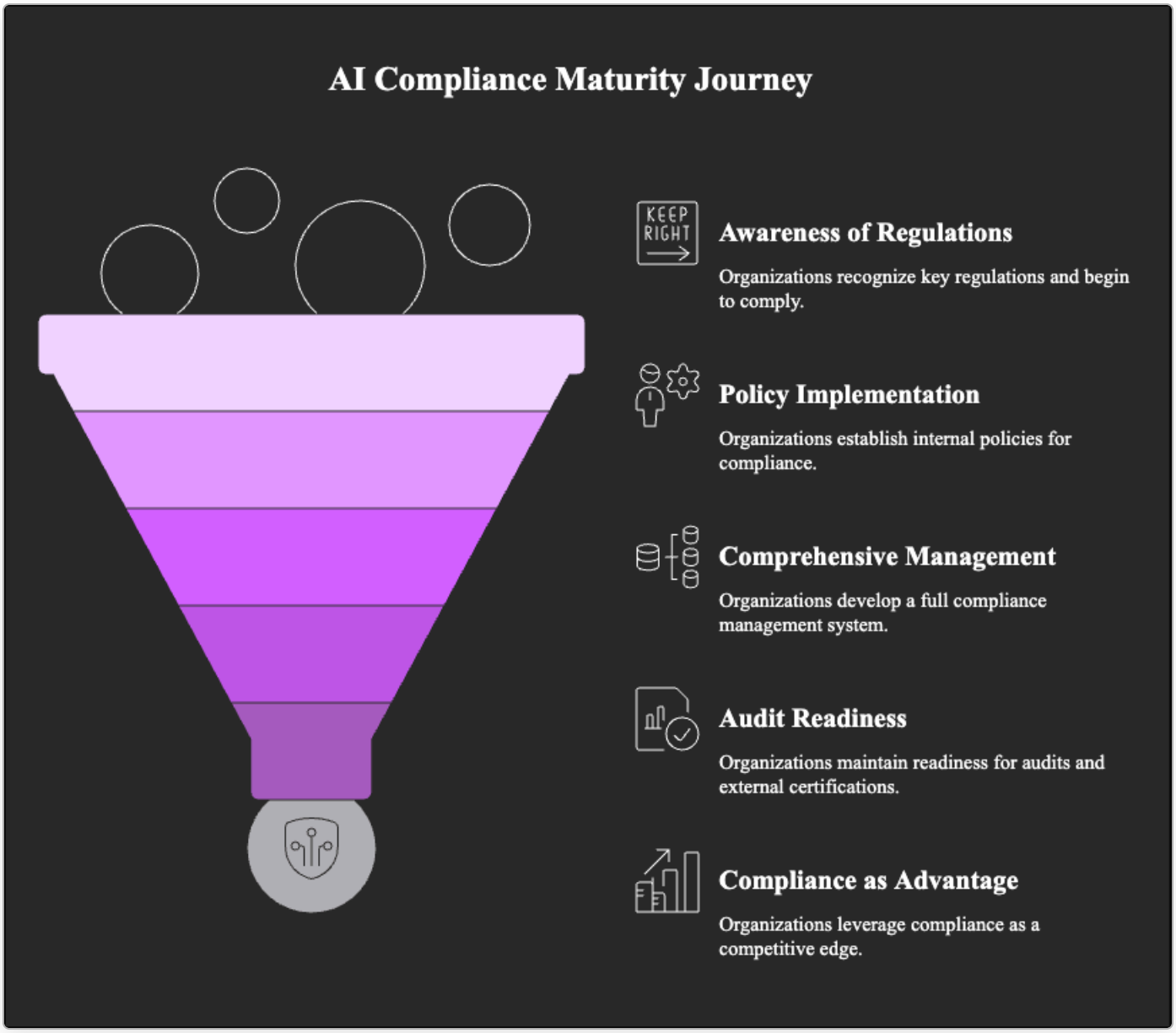


Figure: AI Compliance Program Maturity Stages.

Stage 1: Non-compliant (Ignorant or Defiant) ▶

Stage 2: Aware of Regulations ▶

Stage 3: Implementing Policies and Controls for Compliance ▶

Stage 4: Comprehensive Compliance Management System ▶

Stage 5: Audit Readiness and External Certification ▶

Stage 6: Compliance as Business Enabler ▶

Stage 7: Thought Leader and Shaper in AI Compliance ▶

By assessing themselves against this compliance maturity model, companies can see if they are merely reactive (Stages 1-2), building a program (3-4), solid and proactive (5-6) or truly leading (7). This is particularly crucial for industries like finance, healthcare, or any domain where regulation is heavy – being at higher maturity is not just good practice, it's necessary for business continuity.

Each of these maturity models provides a seven-stage ladder. Not all organizations will need to reach the top in every area; the target maturity may depend on context. For example, a small startup might aim to reach Stage 4 in most areas to satisfy partners and basic ethical duties, whereas a large multinational or critical infrastructure provider should aspire to Stage 6 or 7 in governance, safety, and compliance because the stakes are higher.

Assessment Criteria:

Companies can use the stages above as benchmarks. For each stage, they should ask: do we exhibit these characteristics? For instance, in AI Governance, are our efforts mostly ad-hoc (Stage 1-2) or do we have formal processes (Stage 4)? By scoring themselves, they identify gaps. Key assessment criteria are typically:

- Existence of policies/structures (governance maturity).
- Coverage and consistency of practices (are processes enterprise-wide or siloed).
- Proactivity vs reactivity (e.g. safety and risk: do we anticipate or just respond).
- Stakeholder feedback (trust: do users trust us? compliance: do regulators consider us compliant?).
- Outcomes/metrics (like number of incidents, trend of improvements).

Best Practices and Challenges Recap:

At each stage, we noted best practices and challenges. Common best practices to progress include: gaining leadership buy-in, implementing training and awareness, adopting standards or frameworks to guide improvements, investing in tools for automation (especially to scale governance and monitoring), engaging stakeholders (users, regulators, third parties) for feedback, and fostering a culture that values these aspects (so it's not seen as hindrance but part of quality). Challenges often revolve around resource constraints, potential slowdowns of innovation, need for expertise, and change management (people following new processes).

In implementing maturity models, organizations should also consider dependency between them: e.g. improving Responsible AI maturity might require improving Governance maturity since governance provides the oversight for ethics. In practice, progress in one area often drives progress in others. A holistic approach (perhaps an overall AI Capability Maturity model combining elements of all) can be developed, but dissecting by domain as we did allows focus on specific competencies and specialized teams (compliance team vs engineering team) to take ownership.

Glossary of AI Terms and Definitions

This glossary clarifies key terms. Where definitions are directly taken from ISO/IEC 22989:2022(E), it is explicitly cited[22].

AI agent

automated (3.1.7) entity that senses and responds to its environment and takes actions to achieve its goals

(Source: ISO/IEC 22989:2022, 3.1.1)

AI component

functional element that constructs an AI system (3.1.4)

(Source: ISO/IEC 22989:2022, 3.1.2)

Artificial intelligence (AI)

<discipline> research and development of mechanisms and applications of AI systems (3.1.4)

(Source: ISO/IEC 22989:2022, 3.1.3)

Note 1 to entry: Research and development can take place across any number of fields such as computer science, data science, humanities, mathematics and natural sciences.

Artificial intelligence system (AI system)

engineered system that generates outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives

(Source: ISO/IEC 22989:2022, 3.1.4)

Note 1 to entry: The engineered system can use various techniques and approaches related to artificial intelligence (3.1.3) to develop a model (3.1.23) to represent data, knowledge (3.1.21), processes, etc. which can be used to conduct tasks (3.1.35).

Note 2 to entry: AI systems are designed to operate with varying levels of automation (3.1.7).

Explainability

property of an AI system (3.1.4) to express important factors influencing the AI system (3.1.4) results in a way that humans can understand

(Source: ISO/IEC 22989:2022, 3.5.7)

Note 1 to entry: It is intended to answer the question "Why?" without actually attempting to argue that the course of action that was taken was necessarily optimal.

Machine learning (ML)

process of optimizing model parameters (3.3.8) through computational techniques, such that the model's (3.1.23) behaviour reflects the data or experience

(Source: ISO/IEC 22989:2022, 3.3.5)

Neural network (NN / neural net / artificial neural network)

<artificial intelligence> network of one or more layers of neurons (3.4.9) connected by weighted links with adjustable weights, which takes input data and produces an output

(Source: ISO/IEC 22989:2022, 3.4.8)

Note 1 to entry: Neural networks are a prominent example of the connectionist approach (3.1.10).

Note 2 to entry: Although the design of neural networks was initially inspired by the functioning of biological neurons, most works on neural networks do not follow that inspiration anymore.

Supervised machine learning

machine learning (3.3.5) that makes only use of labelled data during training (3.3.15)

(Source: ISO/IEC 22989:2022, 3.3.12)

Unsupervised machine learning

machine learning (3.3.5) that makes only use of unlabelled data during training (3.3.15)

(Source: ISO/IEC 22989:2022, 3.3.17)

Bias

systematic difference in treatment of certain objects, people or groups in comparison to others

(Source: ISO/IEC 22989:2022, 3.5.4)

Note 1 to entry: Treatment is any kind of action, including perception, observation, representation, prediction (3.1.27) or decision.

Robustness

ability of a system to maintain its level of performance under any circumstances

(Source: ISO/IEC 22989:2022, 3.5.12)

Transparency

<system> property of a system that appropriate information about the system is made available to relevant stakeholders (3.5.13)

(Source: ISO/IEC 22989:2022, 3.5.15)

Note 1 to entry: Appropriate information for system transparency can include aspects such as features, performance, limitations, components, procedures, measures, design goals, design choices and assumptions, data sources and labelling protocols.

Note 2 to entry: Inappropriate disclosure of some aspects of a system can violate security, privacy or confidentiality requirements.

Accountability

state of being accountable (3.5.1)

(Source: ISO/IEC 22989:2022, 3.5.2, referencing ISO/IEC 38500:2015, 2.3)

Note 1 to entry: Accountability relates to an allocated responsibility. The responsibility can be based on regulation or agreement or through assignment as part of delegation.

Note 2 to entry: Accountability involves a person or entity being accountable for something to another person or entity, through particular means and according to particular criteria.

AI Compliance

This denotes the strict adherence of AI systems to all relevant legal, regulatory, and ethical mandates. It is crucial for ensuring the responsible and risk-mitigated design, development, and deployment of AI technologies. AI compliance involves verifying that AI-powered systems do not contravene any laws or regulations and that the data used for training these systems is collected and utilized in a legal and ethical manner. It also guarantees that AI systems are not employed for discriminatory or manipulative purposes and that they respect individual privacy and do not cause harm.

AI Fairness

This principle ensures that AI systems operate without bias, leading to equitable, just, and non-discriminatory outcomes across all their applications. It prioritizes the relatively equal treatment of individuals or groups in the decisions and actions of AI systems, ensuring that these decisions do not disproportionately or negatively impact individuals based on sensitive attributes such as race, gender, or religion.

AI Governance

This encompasses the comprehensive system of policies, controls, and regulations established to ensure that AI is developed, deployed, and managed in an ethical, transparent, and safe manner. It involves bringing together diverse stakeholders from data science, engineering, compliance, legal, and business teams to align AI systems with overarching business, legal, and ethical requirements throughout the entire lifecycle of machine learning models. Effective AI governance applies rules, processes, and responsibilities to maximize the value derived from automated data products while simultaneously mitigating risks and adhering to legal requirements. Ultimately, it directs AI research, development, and application to safeguard safety, promote fairness, and uphold respect for fundamental human rights.

AI Governance Framework

These are structured models that define the fundamental principles, compliance protocols, and risk management strategies necessary to ensure the ethical and transparent development and deployment of AI. These frameworks provide guidance on a wide range of critical topics, including transparency, accountability, fairness, privacy, security, and overall safety. Depending on an organization's specific needs and level of maturity in AI adoption, these frameworks can be implemented in informal ways based on organizational values, through ad hoc development of

specific policies, or via the establishment of a comprehensive and formal governance structure.

AI Risk Assessment

This involves a systematic evaluation of potential risks associated with AI systems. These risks can include bias in algorithms, vulnerabilities to security threats, exposure to regulatory non-compliance, and potential operational failures. The primary goal of an AI risk assessment is to identify and thoroughly map these potential risks and to subsequently develop effective mitigation strategies to address them. For businesses processing consumer personal information, particularly when utilizing automated decision-making technologies, conducting comprehensive risk assessments is a crucial prerequisite.

Responsible AI

This represents a governance-driven approach to the development of AI that places a high priority on fundamental principles such as fairness, transparency, accountability, trust, safety, and overarching ethical integrity. It involves actively steering the responsible development, careful deployment, and ethical use of AI technologies throughout their entire lifecycle. This approach emphasizes the critical role of human oversight and the paramount importance of aligning AI systems with core human values.

AI Audit

This is a formal review and thorough assessment of an AI system conducted to verify that it operates as intended and that it fully complies with all relevant laws, applicable regulations, and established standards. The primary purpose of an AI audit is to identify and map any potential risks associated with the system and to propose effective strategies for mitigating these risks. Regular AI audits are essential for ensuring the continuous adherence of AI systems to evolving regulations and ethical guidelines.

AI Assurance

This encompasses a comprehensive combination of frameworks, established policies, well-defined processes, and robust controls that are implemented to measure, rigorously evaluate, and actively promote the safety, reliability, and overall trustworthiness of AI systems. AI assurance schemes may include various components such as conformity assessments, thorough impact and risk evaluations, independent AI audits, formal certifications, rigorous testing and evaluation protocols, and verification of compliance with pertinent industry standards.

Adversarial Attack

This represents a significant safety and security risk directed at an AI model. It is initiated by malicious actors who deliberately manipulate the model, often by introducing carefully crafted, deceptive input data. These attacks are specifically designed to deceive AI systems, causing them to generate incorrect or unintended predictions or to make faulty decisions. Adversarial attacks exploit inherent vulnerabilities and limitations within machine learning models, particularly those involving deep neural networks. These attacks can be categorized based on the attacker's level of access to the model: white-box attacks occur when the attacker

has complete access to the model's architecture and parameters, while black-box attacks are launched when the attacker can only interact with the model through its inputs and outputs. Common types of adversarial attacks include evasion attacks, where inputs are modified to cause misclassification; data poisoning, which corrupts training data; inference attacks, aimed at revealing sensitive information; and model extraction, where attackers attempt to replicate the model's functionality.

Data Poisoning

This is a specific type of cyberattack where malicious individuals or groups intentionally manipulate or corrupt the training data used to develop artificial intelligence and machine learning models. By injecting incorrect or biased data points into these training datasets, attackers can subtly or drastically alter a model's behavior, potentially leading to data misclassification or a significant reduction in the overall accuracy and effectiveness of the AI system. Data poisoning attacks can be targeted, aiming to manipulate specific outputs of the model, or non-targeted, with the goal of degrading the general robustness and reliability of the model. Various techniques are employed in data poisoning, including direct data injection of fabricated data, the introduction of subtle backdoor triggers within the data, and clean-label attacks where poisoned data appears correctly labeled, making detection particularly challenging.

Model Drift

This phenomenon refers to the gradual degradation of a machine learning model's performance over time. It occurs due to changes in the underlying data patterns or shifts in the relationships between the input variables and the target variable that the model is trying to predict. Model drift, also known as model decay, can lead to

increasingly inaccurate predictions and flawed decision-making based on the model's outputs. Several types of model drift can occur, including concept drift, where the fundamental relationship between inputs and the target changes; data drift (or covariate shift), where the distribution of the input data itself changes; label drift, where the distribution of the target variable shifts; and feature drift, which involves changes in the distribution of individual input features. Proactive monitoring and effective mitigation strategies to address model drift are essential components of a robust AI governance framework.

Algorithm

In the context of AI and machine learning, an algorithm refers to a well-defined procedure or a set of specific instructions and rules designed to perform a particular task or solve a defined problem using a computer.

Federated Learning

This is a machine learning technique that enables multiple independent entities, often referred to as clients, to collaboratively train a shared model without the need to centralize their data. Instead of transferring raw data to a central server, each participating device or organization trains the model locally using its own data and only shares the updated model parameters. This approach prioritizes data privacy and minimization by bringing the model to the data, rather than the other way around.

Differential Privacy

This is a mathematically rigorous definition that provides a robust framework for developing privacy-preserving technologies. It allows organizations to share valuable information about a dataset by describing statistical patterns of groups within the data while ensuring that all personal information about individual data subjects is withheld. Differential privacy works by adding carefully calibrated statistical noise to the results of a query or analysis performed on the dataset. This noise makes it extremely difficult to discern whether any specific individual's data was included in the dataset or to infer any new information about a particular individual based on the analysis. The goal is to ensure that the inclusion or exclusion of any single individual's data does not significantly alter the overall conclusions drawn from the analysis.

AI Ethics

This broad term encompasses a comprehensive set of values, fundamental principles, and practical techniques that employ widely accepted standards of right and wrong to guide moral conduct throughout the entire lifecycle of AI technologies, from their initial development to their eventual use and sale. AI ethics addresses the potential moral, societal, and even legal implications that may arise from the deployment of AI systems, aiming to ensure that these powerful technologies are developed and utilized in ways that are consistent with human values, respect fundamental rights, and promote overall well-being. Key principles often associated with AI ethics include transparency in how AI systems function, fairness in their outcomes, accountability for their actions, the protection of individual privacy, and the explainability of their decisions.

Fairness (Ethical Context)

Within the ethical considerations of AI, fairness refers to the principle of impartial and just treatment or behavior without any unjust favoritism or discrimination in the way AI systems operate. It prioritizes the relatively equal treatment of all individuals or groups affected by an AI system's decisions and actions. Achieving fairness in AI means ensuring that an AI system's decisions and outcomes do not disproportionately or adversely impact individuals based on sensitive attributes such as race, gender, religion, or other protected characteristics.

Interpretability

This refers to the ability to explain or present the reasoning behind a model's decisions and outputs in terms that are easily understandable by humans. Unlike explainability, which often focuses on providing an explanation after a decision has been made, interpretability emphasizes the design of AI models in a way that inherently facilitates understanding through their structure, the features they utilize, or the algorithms they employ. Interpretable models are often domain-specific and require significant expertise in the relevant field to develop effectively.

GDPR (General Data Protection Regulation)

This is a comprehensive data protection and privacy regulation enacted by the European Union. It has a significant impact on the development and application of AI technologies, particularly when these technologies process the personal data of individuals within the EU. The GDPR establishes strict requirements for the lawful processing of personal data, including the need for justifiable grounds for data management, adherence to principles of data minimization and purpose limitation, and the implementation of anonymization and pseudonymization techniques where appropriate. It also grants individuals a range of rights concerning their personal

data, such as the right to access their data, the right to data portability, the right to receive an explanation for decisions made through automated processing, and the right to be forgotten (data erasure). Furthermore, the GDPR places obligations on organizations to ensure accountability, implement data protection by design and by default, and maintain ongoing supervision of compliance.

CCPA (California Consumer Privacy Act)

This is a landmark data privacy law in the state of California, USA, which aims to provide consumers with greater control over their personal information. The CCPA grants consumers various rights, including the right to know what personal information businesses collect about them and how it is used, the right to opt out of the sale or sharing of their personal information, and the right to request the deletion or correction of their data. Notably, the California Privacy Protection Agency (CPPA) has been actively developing and proposing regulations that specifically address the use of artificial intelligence and automated decision-making technologies (ADMT). These proposed rules would require businesses using ADMT for significant decisions to provide consumers with pre-use notices detailing the purpose and operation of the technology, offer mechanisms for consumers to opt out of its use, and explain how the ADMT affects the consumer. Additionally, the proposed regulations mandate that businesses conduct risk assessments before deploying ADMT in certain contexts, particularly when making significant decisions about consumers or engaging in extensive profiling.

EU AI Act

Officially known as the Artificial Intelligence Act, this is a groundbreaking law enacted by the European Union to govern the development and utilization of AI systems within

its member states. The EU AI Act adopts a risk-based approach to regulation, categorizing AI systems into different levels of risk: unacceptable risk (prohibited), high risk (subject to strict obligations), limited risk (requiring specific transparency measures), and minimal or no risk (largely unregulated). The Act imposes various obligations on both providers and deployers of high-risk AI systems, covering aspects such as data governance, technical documentation, transparency requirements, human oversight mechanisms, and robustness and accuracy standards. This comprehensive legislation aims to foster innovation in AI while simultaneously safeguarding fundamental rights and ensuring the safety of individuals and the public.

NIST AI RMF (AI Risk Management Framework)

This is a voluntary framework developed by the National Institute of Standards and Technology (NIST) in the United States. Its purpose is to provide organizations with a structured set of guidelines to effectively assess and manage the diverse risks associated with the implementation and use of artificial intelligence systems. Unlike a legally binding regulation, the NIST AI RMF offers a comprehensive approach to identifying, evaluating, and mitigating AI-related risks across various sectors. The framework provides guidance on a wide array of critical topics, including ensuring transparency in AI systems, establishing clear lines of accountability, promoting fairness and non-discrimination, protecting data privacy and security, and ensuring the overall safety and reliability of AI technologies. By adopting the NIST AI RMF, organizations can enhance their ability to develop and deploy AI responsibly, building trust and fostering innovation in a secure and ethical manner.

AI Compliance (Regulatory Context)

> In a regulatory context, AI compliance refers to the ongoing process of ensuring that all AI-powered systems and applications within an organization adhere to all applicable laws, relevant regulations, established industry standards, and overarching ethical guidelines. This involves a multifaceted approach that includes verifying the legal and ethical use of AI technologies, ensuring the proper handling and security of data used in AI training and deployment, preventing the use of AI for discriminatory or manipulative purposes, safeguarding individual privacy in the context of AI applications, and promoting the responsible and beneficial deployment of AI for society as a whole. Achieving AI compliance typically requires organizations to establish clear internal policies and procedures, develop comprehensive compliance programs, implement robust monitoring systems to track AI usage and performance, and establish effective AI governance frameworks that guide the responsible development and deployment of these powerful technologies.

Looking Forward

This handbook has provided a comprehensive primer on AI Governance, Safety, Trust/Transparency, Responsible AI, and Risk – explaining key legal/regulatory frameworks and technical aspects tailored to the needs of AI practitioners, compliance officers, executives, and policymakers. As organizations navigate the complex landscape of AI, they must align on terminology and goals (hence the glossary) and collaborate across roles: developers building safe and explainable systems, compliance ensuring laws and ethics are met, executives setting direction and culture, and policymakers creating an environment that rewards responsible innovation. By assessing their maturity in various facets of AI governance using the frameworks presented and following best practices at each stage, organizations can steadily improve. The ultimate aim is to harness the power of AI in a way that is responsible, trustworthy, and aligned with societal values, thereby unlocking AI's benefits while managing its risks. Nonetheless, a lot remains to be determined. Companies should use the need to build AI governance functions and capabilities as an opportunity to

upskill, enable, and extend the knowledge of their current employees. Like cybersecurity, governance can involve teams regardless of their roles or job titles.

AI Usage Disclosure: This document was created with assistance from AI tools. The content has been reviewed and edited by a human. For more information on the extent and nature of AI usage, please contact the author.

About the Author

Khullani M. Abdullahi, J.D.

Founder of Techne AI

Khullani M. Abdullahi specializes in AI governance, compliance, and ethics. With a background in law and technology, Khullani helps organizations navigate the complex landscape of AI regulation and responsible implementation.

LinkedIn Website Newsletter

Additional Resources

Organizations and Initiatives

- NIST AI Risk Management Framework
- EU AI Act Resources

- ISO/IEC 42001
- ISO/IEC 23894 (AI Risk Management)
- ISO/IEC 22989 (AI Concepts & Terminology)
- Partnership on AI
- OECD AI Policy Observatory

Tools and Frameworks

- IBM AI Fairness 360
- Microsoft Responsible AI Toolbox
- Model Cards (Google)

References

[1] Osler, Hoskin & Harcourt LLP. "The role of ISO/IEC 42001 in AI governance". Retrieved from https://www.osler.com/en/insights/updates/the-role-of-iso-iec-42001-in-ai-governance/ ↩

[2] Trustible. "Everything you need to know about the NIST AI Risk Management Framework". Retrieved from https://www.trustible.ai/post/nist-ai-rmf-faq ↩

[3] MIAI, Grenoble Alpes. "Tools for Navigating the EU AI Act: (2) Visualisation Pyramid". Retrieved from https://ai-regulation.com/visualisation-pyramid/ ↩

[4] Visier. "What the GDPR Shows Us About the Future of AI Regulation". Retrieved from https://www.visier.com/blog/what-the-gdpr-shows-us-about-the-future-of-ai-regulation/ ↩

[5] Babl AI. "Navigating the New Frontier: How the EU AI Act Will Impact the Conservation and Restoration of Biodiversity and Ecosystems Industry". Retrieved from https://babl.ai/navigating-the-new-frontier-how-the-eu-ai-act-will-impact-the-conservation-and-restoration-of-biodiversity-and-ecosystems-industry/ ↩

[6] NIST. "AI Risk Management Framework". Retrieved from https://www.nist.gov/itl/ai-risk-management-framework ↩

[7] Medill Spiegel Research Center. "Robots and the NIST AI Risk Management Framework". Retrieved from https://spiegel.medill.northwestern.edu/ai-risk-management-framework/ ↩

[8] Thoropass. "Understanding the NIST AI Risk Management Framework: A complete guide". Retrieved from https://thoropass.com/blog/compliance/nist-ai-rmf/ ↩

[9] GDPR Info. "Art. 22 GDPR – Automated individual decision-making, including profiling". Retrieved from https://gdpr-info.eu/art-22-gdpr/ ↩

[10] Cloudflare. "What is the CCPA (California Consumer Privacy Act)?". Retrieved from https://www.cloudflare.com/learning/privacy/what-is-ccpa/ ↩

[11] OECD. "AI principles". Retrieved from https://oecd.ai/en/ai-principles ↩

[12] American National Standards Institute (ANSI). (2024, May 9). "OECD Updates AI Principles". Retrieved from https://ansi.org/standards-news/all-news/2024/05/5-9-24-oecd-updates-ai-principles ↩

[13] TechTarget. "What Is Artificial Intelligence (AI) Governance?". Retrieved from https://www.techtarget.com/searchenterpriseai/definition/AI-governance ↩

[14] CSET Georgetown. (2021). "Key Concepts in AI Safety: Robustness and Adversarial Examples". Retrieved from https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-robustness-and-adversarial-examples/ ↩

[15] Sánchez, I., et al. (2024). "Evolving AI Risk Management: A Maturity Model based on the NIST AI Risk Management Framework". arXiv:2401.15229. Retrieved from https://arxiv.org/abs/2401.15229 ↩

[16] OECD. "Accountability (OECD AI Principle)". Retrieved from https://oecd.ai/en/dashboards/ai-principles/P9 ↩

[17] GSMA. (2024). "The GSMA Responsible AI Maturity Roadmap" [PDF]. Retrieved from https://www.gsma.com/solutions-and-impact/connectivity-for-good/external-affairs/wp-content/uploads/2024/09/GSMA-ai4i_The-GSMA-Responsible-AI-Maturity-Roadmap_v8.pdf ↩

[18] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)* (pp. 220–229). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596 ↩

[19] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for Datasets. *Communications of the ACM*, *64*(12), 86–92. https://doi.org/10.1145/3458723 (Preprint: arXiv:1803.09010 [cs.DB]) ↩

[20] Organisation for Economic Co-operation and Development (OECD). (2019). *Recommendation of the Council on Artificial Intelligence*. OECD/LEGAL/0449. Retrieved from https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449 ↩

[21] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012 ↩

[22] ISO/IEC 22989:2022(E), *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology*. ISO/IEC. ↩

[23] European Commission. "Regulatory framework proposal on artificial intelligence". Retrieved from https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai ↩

[24] Artificial Intelligence Act EU. "Article 99: Penalties". Retrieved from https://artificialintelligenceact.