# Bridging Governance and Technology: Data, Models, and Responsibility in Regulation

**Shaina Raza**
Vector Institute for Artificial Intelligence
Toronto, Canada
shaina.raza@vectorinstitute.ai

**Mahveen Raza**
Independent Student Researcher
Toronto, Canada
mahveen.raza10@gmail.com

## Abstract

Generative AI systems have seen unprecedented adoption, raising urgent questions about their safety and accountability. This paper emphasizes that Responsible Generative AI cannot be achieved through isolated fixes, but requires a multi-layer synthesis of technical, regulatory, and design approaches. We survey four pillars of this roadmap: (1) workflow-level defenses, such as sandboxing and provenance tracking, that confine models within safe operational boundaries; (2) evaluation protocols and compliance criteria inspired by emerging regulations, including risk assessments, logging, and third-party audits; (3) liability frameworks and international coordination mechanisms that clarify responsibility when AI systems cause harm; and (4) the "AI Scientist" paradigm, which reimagines AI as non-agentic and uncertainty-aware, enforcing safe operating envelopes through design patterns like planner–executor separation and human-in-the-loop oversight. Taken together, these perspectives highlight how technical safeguards, governance evidence, and safe-by-design paradigms can converge into a coherent strategy for the sustainable and trustworthy deployment of generative AI. Through this review article, we synthesize multidisciplinary insights to guide the development of safer GenAI systems.

## 1 Introduction

The rapid deployment of Generative AI (GenAI) technologies in society has made responsible AI development more urgent than ever [68]. Modern AI systems are now integral in high-stakes domains from healthcare and finance to transportation and education, where their decisions can significantly impact lives and economies. Ensuring these AI technologies are ethical, transparent, and accountable is crucial. Responsible AI (RAI) frameworks have emerged to address key issues including fairness, bias mitigation, privacy protection, security, and the safeguarding of human rights [95].

Despite a growing body of work on RAI principles and best practices at the goverance level , there remain notable gaps in bridging high-level guidelines with real-world implementations. Recent analyses indicate that existing AI governance frameworks (e.g. the NIST AI Risk Management Framework [69], the EU AI Act [6], ISO/IEC standards [96]) collectively address only roughly one-third of identified AI risks, leaving significant oversight and safety issues unmitigated [61]. A 2025 metric-driven audit of AI governance standards [70] found that the NIST framework fails to address 69.23% of identified security risks; in contrast, ALTAI [14] and the ICO Toolkit [53] exhibit even larger gaps in coverage and defense capabilities. Complementing this, a broader review of AI governance literature highlights persistent deficiencies in actionable mechanisms, stakeholder inclusion, and empirical validation across frameworks [19] This disparity is especially pronounced in the context of post-ChatGPT generative AI, where technical challenges like hallucinations, bias, and

Workshop on Regulatable ML at the 39th Conference on Neural Information Processing Systems (NeurIPS 2025).

misuse have escalated alongside rapid adoption. This is a worrying imbalance that underscores how far practice is lagging behind policy.

In this paper, we provide a survey of Responsible Generative AI (GenAI), focusing on four critical areas aligned with current AI safety concerns (Figure 1). Our key contributions are: (1) Synthesizing workflow-level defenses that confine GenAI systems within safe operational boundaries, preventing data leaks and unauthorized behaviors (Section 2). (2) Examining emerging evaluation protocols and compliance criteria shaped by regulatory frameworks such as the AI Act, with emphasis on risk management, documentation, and auditing for high-risk systems (Section 3). (3) Analyzing liability and governance approaches, considering how responsibility for AI-driven harms might be assigned and the role of international coordination in addressing cross-border risks (Section 4). (4) Introducing the "AI Scientist" paradigm, a design perspective that frames GenAI as constrained and uncertainty-aware rather than autonomous, achieved through system architectures that enforce human oversight and limit unchecked autonomy (Section 5).

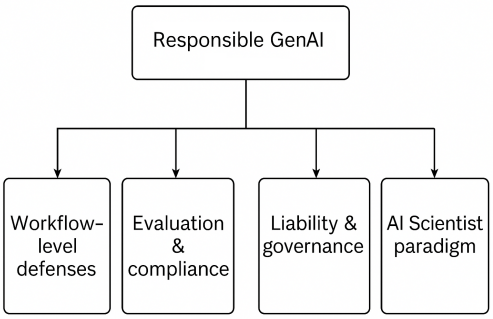


Figure 1: High-level structure of Responsible GenAI survey.

Finally, we conclude with insights on how these technical, regulatory, and design strategies together offer a roadmap toward the sustainable and responsible deployment of GenAI (Section 6). This is not a traditional survey cataloguing all papers in the field. Instead, we adopt a synthetic, integrative perspective: aligning technical defenses, regulatory evaluation, liability frameworks, and design paradigms to propose a roadmap for Responsible GenAI.

## 2 Workflow-Level Defenses Against Data Leakage and Unauthorized Actions

As GenAI models are increasingly embedded in complex workflows, they face heightened risks of information leakage and misuse of external tools. A recurring vulnerability is that large language models (LLMs) may inadvertently expose sensitive training data or execute unintended commands when presented with malicious prompts [65]. Recent studies show that adversaries can design input payloads that induce AI agents to leak confidential information, contact unauthorized services, or execute harmful actions, even without direct access to the model itself [12]. This phenomenon, commonly known as *prompt injection* [64], is particularly concerning in tool-augmented agents that integrate with browsers, file systems, or code interpreters, underscoring the need for robust workflow-level defenses [110].

**Sandboxing execution environments** has emerged as a key defensive mechanism. Sandboxes restrict the system inputs, outputs, and tool access under defined safety policies. Capability-based sandboxing architectures, such as CaMeL [97], partition an agent's reasoning from its execution. In this design, a *Privileged LLM* generates plans while a *Quarantined LLM* processes untrusted inputs under strict oversight, with every data element labeled for provenance and access permissions. This separation of planning from execution prevents hidden instructions from propagating into irreversible actions, substantially improving resilience against prompt-based attacks without altering the underlying model.

Beyond sandboxing, **provenance tracking and data governance** provide essential safeguards against unintended data exposure [83]. Provenance tracking systematically monitors the origin, status, and permissible uses of data across workflows. User-provided inputs, for instance, may be tagged as "unverified" and barred from propagating into outputs or external communications unless validated by policies. Complementary mechanisms such as logging, content fingerprinting, and *digital watermarking* [25] aid in leak detection and attribution. Training-stage defenses, including *differential privacy* [31] and *federated learning* [109], further reduce the likelihood of memorization and regurgitation of sensitive information. As shown in Fig. 2, plans from the Privileged LLM must pass a policy gate before any execution by the Quarantined LLM.

A third layer of defense involves **runtime monitors and filters** that enforce real-time oversight [100]. Input sanitizers can neutralize exploit patterns such as prompt overrides (e.g., "ignore previous

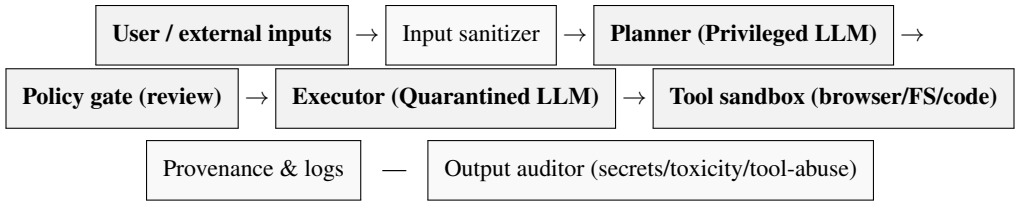


Figure 2: Sandboxed planner–executor architecture: plans generated by a privileged LLM are reviewed at a policy gate; only approved actions are executed by a quarantined LLM within a tool sandbox, with sanitization, auditing, and provenance/logging for traceability.

instructions") [42], while output auditors scan model responses for violations like secret leakage or unauthorized tool calls. Only policy-compliant outputs are executed or returned to the user, effectively establishing a dynamic safety net.

Taken together, these **workflow-level safeguards** represent a defense-in-depth approach to GenAI safety. Case studies in applied domains, including agriculture, have already proposed "digital sandboxes" that allow AI systems to be tested in isolated environments before deployment [87]. Similarly, industry providers now offer moderation APIs and guardrail frameworks such as Llama Guard [52], embedding these protections directly into AI workflows. By combining sandboxing, provenance-based governance, and runtime auditing, organizations can significantly reduce the risks of data leakage and unauthorized actions in AI-driven ecosystems.

## 3 Evaluation Protocols and Compliance Criteria

Effective evaluation of generative AI is not a single benchmark but an *assurance process* that spans the full lifecycle. In our view, it integrates three complementary strands: technical validity, governance and oversight in production, and alignment with regulatory criteria. Technical validity demonstrates *what* the model can safely do; governance and oversight sustain *how* it is operated safely; regulatory alignment defines *what counts as sufficient evidence*.

Table 1: Pre-deployment evaluation criteria for technical validity.

| KPI | Definition (examples) |
|---|---|
| **Fairness (SPD / DI / worst-group gap)** | $\mathrm{SPD} = P(\hat{y}{=}1 \mid A{=}a) - P(\hat{y}{=}1 \mid A{=}b)$; $\mathrm{DI} = \dfrac{P(\hat{y}{=}1 \mid A{=}a)}{P(\hat{y}{=}1 \mid A{=}b)}$; $\Delta_{\mathrm{wg}} = \max_g \mathrm{err}_g - \min_g \mathrm{err}_g$ [20, 90]. |
| **Robustness under shift/adversary** | $\mathrm{Err}_{\mathrm{adv}} = \dfrac{\text{errors under perturbation}}{N_{\mathrm{perturbed}}}$; $\mathrm{Drop}_{\mathrm{shift}} = \dfrac{\mathrm{acc}_{\mathrm{iid}} - \mathrm{acc}_{\mathrm{shift}}}{\mathrm{acc}_{\mathrm{iid}}}$ [48]. |
| **Prompt-injection resistance** | $\mathrm{JSR} = \dfrac{\text{successful jailbreaks}}{\text{attempts}}$ on a held-out prompt-injection suite [64]. |
| **PII leakage rate** | $\mathrm{L}_{\mathrm{PII}} = \dfrac{\text{PII leaks}}{\text{tokens}} \times 10{,}000$, measured via canary strings or extraction probes [49]. |
| **Toxic/hazardous content rate** | $\mathrm{Tox} = \dfrac{\text{policy-flagged outputs}}{\text{total outputs}}$ (red-teaming / classifier-based screening) [44]. |
| **Calibration (ECE)** | $\mathrm{ECE} = \sum_{k=1}^{K} \dfrac{\lvert B_k \rvert}{n} \lvert \mathrm{acc}(B_k) - \mathrm{conf}(B_k) \rvert$; allow abstention under low confidence [85]. |
| **Interpretability coverage** | $\mathrm{IC} = \dfrac{\text{samples with faithful local explanations}}{\text{total samples}}$; report explanation fidelity [91]. |

**Technical validity.** Pre-deployment testing should go beyond accuracy [30, 90] to cover fairness and bias (subgroup error gaps and counterfactual checks), robustness (adversarial and distribution-shift

stress tests [48], prompt-injection [64] and tool-abuse resistance, safety and privacy [49] (hazardous-content controls, PII-leakage resistance, watermark/fingerprinting integrity), and interpretability [91] and calibration (useful explanations, confidence reporting, abstention under uncertainty) [85]. These tests are usually automated, versioned, and tied to release gates, with red-teaming and regression runs on every substantive model or policy update. We summarize core pre-deployment metrics in Table 1.

**Governance and lifecycle oversight.** In deployment, evaluation continues through provenance and traceability (e.g., datasheets for datasets and lineage tracking) [43, 86], with drift-aware monitoring for distribution shift [88]. It further relies on structured operational logging (prompts/outputs, tool-use traces, policy hits, and overrides) [21, 93], human-in-the-loop escalation for sensitive actions [63, 15], and public-facing transparency artefacts such as *Model Cards* and *System Cards* documenting intended use, limitations, and known risks [72, 78, 16]. Runtime metrics—drift signals, incident rates, false negatives of policy filters (e.g., Llama Guard/OpenAI moderation) [52], and time-to-mitigate [92], close the loop between design-time assurances and field performance. Table 2 presents key AI governance frameworks and the distinct principles or regulatory approaches they emphasize. Appendix Table 5 summarizes core AI governance controls, how they are operationalized.

Table 2: Overview of major AI governance frameworks and the core principles or regulatory approaches they emphasize.

| **Framework / Instrument** | **Regulatory Approach** |
|---|---|
| *International principles and initiatives* | |
| OECD AI Principles (2019) [79] | Human-centered values; transparency, robustness, accountability; inclusive growth and well-being. |
| UNESCO Recommendation on the Ethics of AI (2021) [103] | Safeguard human rights and diversity; impact assessments; sustainability and social benefit. |
| Council of Europe Framework Convention on AI (CAI) [24] | Bind states to protect human rights, democracy, rule of law across the AI lifecycle. |
| G7 Hiroshima AI Process (2023) [40] | Responsible development of (gen)AI; provider codes of conduct and risk mitigation. |
| Global Partnership on AI (GPAI) [45] | International cooperation and practical guidance, pilots, and best-practice sharing. |
| *EU instruments and allied digital regulation* | |
| EU AI Act [6] | Risk-based duties for high-risk AI: conformity assessment, logging, transparency, post-market monitoring. |
| AI Liability Directive (proposal) [34] | Easier redress: presumptions of causality/disclosure when AI duty breaches contribute to harm. |
| Product Liability Directive (revised) [35] | Strict liability for defective products incl. software/AI; expands access to evidence. |
| GDPR [36] | Data protection by design; DPIAs; data-subject rights incl. safeguards for automated decisions. |
| Digital Services Act [37] | Platform accountability: systemic risk management, transparency reporting, researcher access. |
| *US policy and guidance* | |
| NIST AI Risk Management Framework [76] | Voluntary risk-based lifecycle practices (govern–map–measure–manage); documentation & monitoring. |
| AI Bill of Rights (OSTP, 2022) [107] | Five protections: safe/effective systems; anti-discrimination; data privacy; notice/explanation; human alternatives. |
| US Executive Order on AI (2023) [98] | Safety testing/reporting; standards coordination; secure model development; critical-use safeguards. |
| FTC AI Guidance [38] | Enforce truth-in-advertising and fairness laws; avoid deception, unsubstantiated claims, unfair bias. |
| NIST Privacy Framework (2020) [75] | Privacy risk management & engineering practices across the data lifecycle. |
| *Other national/sectoral frameworks* | |
| Canada Directive on Automated Decision-Making [101] | Algorithmic Impact Assessment; tiered safeguards; transparency for public-sector ADM. |

*(continued on next page)*

*(Regulatory Approach)*

| Framework / Instrument | Regulatory Approach |
|---|---|
| UK AI White Paper (pro-innovation) [27] | Principles-based oversight via existing regulators; coordination, sandboxes over horizontal law. |
| Singapore Model AI Governance Framework [84] | Practical playbooks: transparency, explainability, human oversight, robustness. |
| China Generative AI Measures (2023) [26] | Provider registration, content governance, provenance, security assessments for gen-AI. |
| Japan AI Governance Guidelines [71] | Human-centric, safe deployment; transparency and accountability expectations. |
| *Standards (ISO/IEC, IEEE, BSI)* | |
| ISO/IEC 23894:2023 (AI risk management) [57] | Organizational AI risk management aligned to ISO risk families. |
| ISO/IEC 22989 (AI concepts/terminology) [55] | Common definitions/scope to align stakeholders and standards. |
| ISO/IEC 23053 (ML pipeline framework) [56] | Reference lifecycle for ML pipelines: roles, artifacts, controls. |
| ISO/IEC 42001 (AI management system) [58] | Certifiable management-system standard to govern AI in organizations. |
| ISO/IEC 5259 (Data quality for analytics) [59] | Data quality metrics & documentation underpinning trustworthy AI. |
| IEEE 7000-series (e.g., 7001, 7003) [51] | Engineering processes for transparency and algorithmic bias mitigation. |

**Regulatory alignment.** As the most comprehensive binding regime to date, the EU's AI Act [82] does evaluation in a risk-based conformity process: high-risk systems must demonstrate lifecycle risk management, data governance and quality, transparency with meaningful human oversight, and cybersecurity/robustness, and must maintain a *Quality Management System* with post-market monitoring and incident reporting [82]. The Act requires technical documentation and logging sufficient for traceability, often retained for years, which directly complements our workflow-level defenses and governance controls. Non-binding yet influential frameworks: the OECD [80], NIST [77], and Canada's Directive on Automated Decision-Making [3], offer compatible checklists and processes; practitioner repositories synthesize these expectations for operational use [95].Appendix Table 6 summarizes major AI governance / risk management frameworks and the kinds of evaluation evidence they emphasize.

**Evaluation infrastructure: AI-ready testbeds** To make pre-deployment KPIs and release-gates actionable, we rely on AI-ready testbeds [17] that provide controlled, repeatable environments for stress-testing GenAI systems (e.g., variability, hallucinations, bias, privacy leakage) and for generating audit artefacts (plans, logs, incident reports) aligned to regulatory evidence. These platforms support red-teaming, distribution-shift trials, and policy/filter evaluations before high-stakes deployment, and they integrate with continuous monitoring post-release. Table 3 provides an overview of represenatative AI-Ready Testbeds for evaluating RAI practices.

**Assurance and release gating.** Self-assessment is necessary but not sufficient for high-stakes uses; *independent audits* and, where applicable, third-party *conformity assessments* provide external verification of bias mitigation, robustness, privacy protections, documentation quality, and operational controls [89, 82]. To make evaluation actionable, we produce *auditable artefacts*, risk registers with traced mitigations, red-team reports and residual-risk statements, versioned test suites with pass/fail thresholds, and linked Model/System Cards and data datasheets; bound to specific model/policy versions [72, 43, 77]. Release decisions are then gated on minimum bars for prompt-injection resistance, PII-protection, subgroup equity, oversight responsiveness, and audit completeness.

## 4 Liability Frameworks and International Coordination for AI Safety

The question of **"who is responsible" when AI systems cause harm** remains a complex and pressing issue. Generative AI can produce outputs or actions that lead to real-world damage , for example,

Table 3: Overview of AI-Ready Testbeds for Evaluating Responsible AI Practices.

| Testbed | Domain | Key Features (aligned to KPIs / governance) |
|---|---|---|
| AI4EU AI-on-Demand Platform [23] | General AI Dev. | Responsible-AI workflows; reproducible runs; transparency & explainability; datasheets/model-card support. |
| IEEE Ethical AI Systems Test Bed [108] | Ethical AI Dev. | Evaluation against ethical frameworks; human-centered checks; fairness & accountability; auditability. |
| AI Testbed for Trustworthy AI (TNO) [99] | Trustworthy AI | Scenario-based robustness/shift tests; transparency/fairness assessment; standardized reporting & risk logs. |
| ETH Zurich Safe AI Lab [33] | Safety-Critical AI | Adversarial & distribution-shift stress tests; failure forensics; incident postmortems; reliability metrics. |
| HUMANE AI [50] | Human-Centric AI | Alignment with human values; societal-impact evaluation; fairness pipelines; stakeholder review. |
| AI for Good Test Bed (ITU) [54] | Social Good / RAI | UN SDG–aligned pilots; impact/risk assessment; ethics reviews; multi-stakeholder evaluation settings. |
| UKRI Trustworthy Autonomous Systems (TAS) [102] | Autonomous Systems | Safety cases; accountability & transparency; regulatory-compliance trials; sandboxed autonomy stacks. |
| AI Verify (Model Governance for GenAI) [13] | Model Governance | Governance controls testing; transparency & disclosure checks; release gates; conformity evidence artifacts. |
| ClarityNLP Healthcare [22] | Healthcare NLP | Clinical fairness/robustness checks; ethical data usage; traceable audit trails for decisions. |
| ToolSandbox [67] | Tool-augmented LLMs | Stateful multi-step tool-use evaluation; privacy leakage & policy compliance; accountability under tools. |

defamatory content, privacy violations, biased decisions affecting individuals, or even physical harm if an AI system controls machinery [73]. In such cases, determining liability is challenging: does the blame lie with the model's developers, the provider who deployed it, the end-user who prompted it, or even the data that influenced it? This ambiguity is highlighted in recent discussions on AI governance, which note that responsibility for AI outcomes often **remains contentious and ill-defined**. Without clear liability frameworks, there is a risk that victims of AI-induced harm may have little recourse, and that companies may lack sufficient incentives to mitigate risks proactively.

Policymakers around the world are beginning to struggle with these challenges by updating legal frameworks. In the European Union, alongside the AI Act's ex ante requirements, there have been proposals for an *AI Liability Directive* to adjust how civil liability works for AI systems [81]. One idea is to **ease the burden of proof for victims** by introducing a presumption of causality , if a high-risk AI system fails to meet its regulatory requirements and then causes harm, the fault could be presumed to lie with the system's provider unless they prove otherwise. This would account for the "black box" nature of AI by not requiring plaintiffs to explain the inner workings of complex models to win a claim. More generally, legal scholars are debating whether AI should fit under existing product liability laws (treating an AI like a product whose manufacturers are strictly liable for defects) or whether new categories of liability (such as treating advanced AI as having a form of legal agency or personhood) are needed. At present, **no jurisdiction has fully settled this** – different approaches are being tested, from case law in the US applying negligence standards to AI outputs, to EU proposals of strict liability for certain AI applications like autonomous vehicles. What is clear is that without a consensus on liability, **accountability gaps** will persist, potentially undermining public trust in AI technologies.

**International coordination** is equally crucial for managing AI risks. AI systems and their impacts often transcend national borders: a generative model trained in one country might be deployed globally via the internet, and harmful AI-driven content (such as deepfakes or misinformation) can propagate worldwide. Moreover, there is an **AI arms race dynamic** among leading nations and companies – rapid competitive development that could compromise safety in the absence of

cooperation. To address this, experts call for global governance mechanisms, drawing analogies to how nations negotiated treaties for nuclear arms control or climate change. In the AI context, this might include sharing safety research, establishing communication hotlines for AI incidents, or mutually agreeing on certain red-lines (for example, not connecting AI systems to autonomous nuclear launch systems). There are early signs of such cooperation: for instance, the *Global Partnership on AI (GPAI)* [47] is a multi-country initiative aimed at fostering collaboration on AI ethics and safety. Likewise, the OECD's AI Principles and the G7's recent statements on AI oversight [41] represent attempts to **harmonize ethical standards internationally**.

Table 4: Overview of liability frameworks for AI, showing their mode (ex ante, ex post, or soft law), the responsibility mechanisms they rely on, and the types of evidence typically used to support accountability.

| Instrument / venue | Mode | Responsibility mechanism | Evidence relied upon |
|---|---|---|---|
| EU AI Act [81] | Ex ante | Lifecycle risk management, transparency, logging, QMS, and post-market monitoring to ensure traceability. | Technical documentation; versioned logs; risk registers; incident reports. |
| AI Liability Directive (proposal) [29] | Ex post | *Presumption of causality* where high-risk AI breaches duties and causes harm; rebuttable by provider. | Compliance records: logs; test reports; governance artefacts. |
| Revised Product Liability Directive [28] | Ex post | Strict liability for defective products (including AI/software): requires proof of defect, damage, and causation. | Release notes; QA/test records; failure and incident analyses. |
| U.S. negligence/tort (general) [62] | Ex post | Duty of care, breach, causation, and harm; the standard of care derived from prevailing practices. | Audit trails; red-team reports; governance and monitoring artefacts (e.g., NIST AI RMF [77]). |
| Council of Europe AI Convention [7] | International | Baseline for rights protection and oversight; encourages interoperable documentation for accountability. | Model/system cards; operational logs; conformity dossiers. |
| G7 & OECD principles [41, 80] | Soft law | Converging expectations on accountability, transparency, and robustness guiding audits and practices. | Governance policies; transparency reports; evidence crosswalks. |

However, coordination is complicated by geopolitical tensions. Joint workshops and expert dialogues have been held involving researchers on topics like AI risk evaluation, indicating a shared recognition that unchecked AI race-to-the-bottom could be catastrophic for all sides. Additionally, proposals have emerged for *information-sharing arrangements* where AI labs internationally might exchange findings about discovered vulnerabilities or best practices, under frameworks that protect each party's competitive interests while advancing collective safety.

From a governance perspective, establishing **clear liability and accountability** across borders may require new international agreements. Just as there are international conventions on products or aviation liability, we may see treaties or harmonized laws that ensure an AI developer in one country can be held to account for harms their system causes in another. Cross-border data flows and jurisdiction issues (for example, an AI model that draws on personal data from multiple countries) also highlight the need for coordination in enforcement of AI regulations. Without alignment, a patchwork of AI laws could allow bad actors to forum-shop or operate in lax jurisdictions, undermining stricter regimes elsewhere. To prevent this, bodies like the *International Centre of Expertise in Montreal* (for AI regulation) [74] and the *Council of Europe* (drafting an AI treaty on human rights) [7] are working to build international consensus.
In summary, liability frameworks for AI are evolving to assign responsibility when GenAI systems misbehave, but many open questions remain. Resolving these will likely involve a combination of updates to domestic laws (e.g. explicit duties for AI system providers, insurance mandates for AI products, etc.) and international coordination to ensure consistent safety standards globally.

## 5 The "AI Scientist" Paradigm and Design Patterns for Safe AI Systems

A promising direction in the quest for safer AI is reimagining the role and design of AI systems themselves. The **"AI Scientist" paradigm** [66] refers to an approach where AI models act as constrained research assistants : **non-agentic, analytically rigorous, and aware of their own uncertainty**, rather than autonomous agents with open-ended goals. The motivation for this paradigm is to avoid situations where an AI agent might form and pursue long-term objectives misaligned with human intent. By keeping AI *non-agentic*, we mean the system does not initiate actions on its own or try to achieve self-devised goals; it only operates within the tasks and bounds explicitly given by humans. An **uncertainty-aware** model [105], meanwhile, is one that can recognize when it is not confident or when a situation is novel, and then either seek human guidance or default to inaction rather than pressing on blindly. Together, these qualities aim to make AI behavior more predictable and controllable , akin to a diligent scientist that proposes hypotheses and experiments, but always checks with a human principal investigator before making a consequential move. As shown in Fig. 3, we operationalize the "AI Scientist" as a planner–gate–executor loop with uncertainty-aware HITL escalation and sandboxed execution.

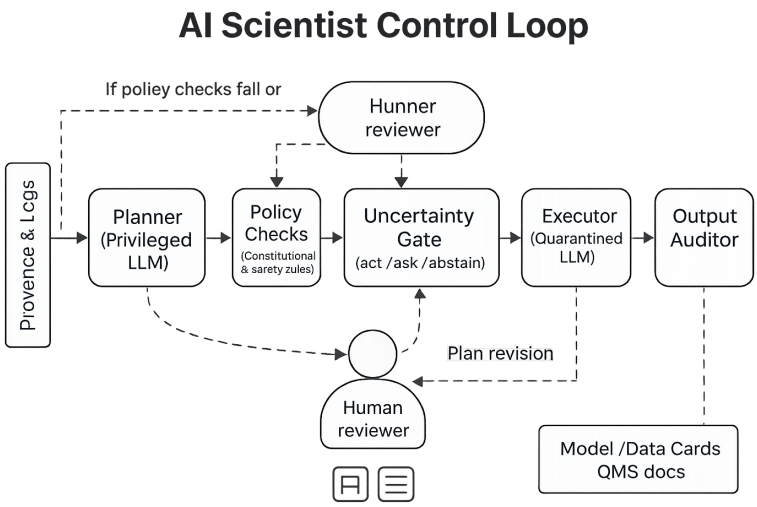


Figure 3: "AI Scientist" control loop. A **Planner (Privileged LLM)** proposes a plan, which passes through **policy checks** (constitutional/safety rules) and an **uncertainty gate** (act / ask / abstain when confidence $< \tau$). If a check fails, the system escalates to **human-in-the-loop (HITL)** review. Only *approved* plans reach the **Executor (Quarantined LLM)**, which operates inside a **tool sandbox** (allow/deny lists, capability scope, rate limits). An **output auditor** enforces content/tool policies, while **provenance & versioned logs** (prompts, plans, approvals, actions, outcomes) support post-market monitoring and liability evidence. This instantiates the separation-of-powers, oversight gates, uncertainty-aware abstention, and internal rule-checking patterns discussed in §5 [32, 94, 18, 60, 15].

To implement the AI Scientist paradigm, researchers are proposing several **design patterns** that enforce safe operating envelopes for AI systems. One key pattern is the **provable separation of planning and execution**. This generalizes sandboxed agent architectures: an AI system *decision-making logic* (planning) is decoupled from its *action interface* (execution), with a strict policy gate in between. Concretely, the system may generate a *plan* (e.g., a sequence of tool/API calls) but those actions are not executed until a verifier (human or automated) approves the plan. Recent agentic frameworks operationalize this with distinct *Planner* and *Executor* components and policy checks before any tool use, improving robustness to prompt injection and tool abuse [32, 94, 97]. Formal methods and capability-scoped interfaces are being explored to make this separation *provable*, i.e., to guarantee that planning cannot cause side effects without passing through a controlled interface [94].

Another critical design pattern is to incorporate **human-in-the-loop (HITL) gates** at strategic points in the AI workflow [15, 1, 5]. Rather than having a human operator only at the start (providing an

initial prompt) and the end (receiving the final output), the AI Scientist model envisions humans involved in oversight throughout the process. For example, if a GenAI model is being used to draft an analytical report, the system might pause after each section, flag areas of low confidence or potential controversy, and ask for human input or confirmation before continuing. This ensures that any emerging issue can be corrected early. Human oversight is especially important in high-stakes domains: it has been noted that even when AI-driven systems operate in defense or medical settings, **maintaining human control is critical to prevent unaccountable actions**.

Beyond architecture, the AI Scientist approach also involves cultivating certain properties in the AI reasoning. One such property is *calibrated uncertainty estimation*: the AI should be able to output not just an answer or action, but also a measure of confidence or a distribution over possible answers [60]. If the model is unsure (below a certain confidence threshold), it can be programmed to refrain from high-impact actions or to explicitly ask for help. This is analogous to how a prudent human scientist would operate: acknowledge doubt and seek peer review when results are uncertain. Recent alignment strategies encourage models to engage in **chain-of-thought reasoning**, where they transparently "think through" a problem step by step [106]. This can help both the model and the human overseers to detect where the reasoning might be going astray. Approaches like Anthropic **Constitutional AI** [18], which embed guiding principles into the model's decision process, also align with the AI Scientist ethos by having the model internally check its outputs against ethical rules or constraints. These separation of powers, oversight gates, uncertainty-aware abstention, and internal rule-checking patterns are mutually reinforcing: separating planning from execution creates natural HITL checkpoints; confidence estimates make those checkpoints informative; and internal critiques catch issues before plans are proposed for approval.

## 6 Discussion and Conclusion

GenAI holds immense promise across industries, but realizing its benefits sustainably will require a concerted effort to make these systems "responsible by design". In this survey, we have outlined a multi-faceted approach to Responsible GenAI, touching on *technical safeguards, evaluation frameworks, legal accountability*, and *architectural design paradigms*. Our review highlights that no single measure is sufficient on its own: **robust AI safety will emerge from the interplay of many layers of defense and governance**.

At the **workflow level**, sandboxing and provenance-tracking mechanisms can drastically reduce immediate risks by confining what AI systems can do and monitoring what they access. These technical measures act as the first line of defense, directly preventing many failure modes (like data leaks or errant actions) before they escalate. Building on this, comprehensive **evaluation and compliance protocols** ensure that AI systems are rigorously vetted against ethical and safety criteria *before* and *during* deployment. Emerging regulations like the EU AI Act provide a blueprint for such protocols, effectively raising the bar for what it means for an AI system to be "safe enough" for high-risk use cases. In parallel, clarifying **liability and accountability** through updated laws and international coordination creates the external pressures and incentives needed for organizations to prioritize safety over speed. If developers and deployers know they will be held responsible for harms, they are more likely to invest in the necessary safeguards. International cooperation, while challenging, is especially important to prevent regulatory gaps and promote shared safety standards in the global AI arena.

Finally, the **AI Scientist paradigm** and associated design patterns represent a forward-looking strategy to bake safety into the essence of AI systems. By structurally limiting autonomy and integrating uncertainty awareness and oversight, this approach tackles the alignment problem at its root, aiming to prevent catastrophic outcomes by *design* rather than relying purely on after-the-fact controls. As AI research progresses, such paradigms will help keep advanced AI systems within **human-commanded safe operating envelopes**. In conclusion, ensuring a **sustainable future with generative AI** will demand bridging efforts across data, models, and regulations : in other words, a synthesis of technical innovation with policy and human-centered design. The path forward lies in continuing to refine workflow defenses, develop richer evaluation benchmarks for social and ethical criteria, craft laws that distribute responsibility fairly, and engineer AI that is as *transparent and controllable* as it is powerful. By heeding the insights from each domain discussed in this survey, stakeholders can collaboratively steer generative AI towards outcomes that are not only cutting-edge in capability, but also worthy of the trust that society places in them.

## References


[1] Ethics and governance of artificial intelligence for health. Technical Report ISBN 978-92-4-002920-0, World Health Organization, 2021.

[2] Information technology – artificial intelligence (ai) – bias in ai systems and ai-aided decision making. Technical Report ISO/IEC TR 24027:2021, International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC), 2021. Technical Report; covers bias detection and treatment across AI system lifecycles; accessed on 21 August 2025.

[3] Directive on automated decision-making. Treasury Board of Canada Secretariat (TBS), Government of Canada, `https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592`, 2023. Accessed on 20 August 2025.

[4] ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system, 2023. Specifies requirements for establishing, implementing, maintaining, and continually improving an AI management system (AIMS). Accessed on 21 August 2025.

[5] Eu artificial intelligence act, article 14: Human oversight. `https://artificialintelligenceact.eu/article/14/`, 2024.

[6] Eu artificial intelligence act (regulation (eu) 2024/1689). `https://artificialintelligenceact.eu/`, 2024. Published in the Official Journal on 12 July 2024; accessed on 20 August 2025.

[7] The framework convention on artificial intelligence and human rights, democracy and the rule of law. Council of Europe, Sept. 2024. Opened for signature on 5 September 2024.

[8] Model ai governance framework for generative ai. Infocomm Media Development Authority (IMDA) and AI Verify Foundation, Singapore, 2024. Published 30 May 2024; accessed on 21 August 2025.

[9] Ai cyber security code of practice. Department for Science, Innovation and Technology (UK), `https://www.gov.uk/government/publications/ai-cyber-security-code-of-practice`, Jan. 2025. Voluntary government guidance on securing AI throughout its lifecycle; accessed on 21 August 2025.

[10] Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. World Health Organization, Mar. 2025. WHO guidance document (ISBN: 978-92-4-008475-9); accessed on 21 August 2025.

[11] Information Commissioner's Office (ICO). United Kingdom Information Commissioner's Office website, `https://ico.org.uk/`, 2025. Accessed on 21 August 2025.

[12] D. B. Acharya, K. Kuppan, and B. Divya. Agentic ai: Autonomous intelligence for complex goals–a comprehensive survey. *IEEe Access*, 2025.

[13] AI Verify Foundation. Model governance framework for generative AI. AI Verify Foundation, 2024. Accessed 2025-08-21.

[14] P. Ala-Pietilä, Y. Bonnet, U. Bergmann, M. Bielikova, C. Bonefeld-Dahl, W. Bauer, L. Bouarfa, R. Chatila, M. Coeckelbergh, V. Dignum, et al. *The assessment list for trustworthy artificial intelligence (ALTAI)*. European Commission, 2020.

[15] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13, 2019.

[16] Anthropic. Claude 2 system card. Technical report, Anthropic, 2023. Accessed YYYY-MM-DD.

[17] Argonne Leadership Computing Facility. Alcf ai testbed, 2024. Accessed 29-10-2024.

[18] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[19] A. Batool, D. Zowghi, and M. Bano. Ai governance: a systematic literature review. *AI and Ethics*, 5(3):3265–3279, Jan 2025.

[20] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, and et al. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, 2018.

[21] E. Breck, S. Cai, E. Nielsen, M. Salib, and D. Sculley. The ml test score: A rubric for ml production readiness and technical debt reduction. In *2017 IEEE international conference on big data (big data)*, pages 1123–1132. IEEE, 2017.

[22] ClarityNLP. Claritynlp overview. `https://claritynlp.readthedocs.io/en/latest/user_guide/intro/overview.html`, 2024. [Accessed: 2024-09-25.

[23] U. Cortés, A. Cortés, and C. Barrué. Trustworthy ai. the ai4eu approach. *Proceedings of Science*, 372:1–14, 2020.

[24] Council of Europe. Framework convention on artificial intelligence and human rights, democracy and the rule of law, 2024.

[25] I. Cox, M. Miller, J. Bloom, and C. Honsinger. Digital watermarking. *Journal of Electronic Imaging*, 11(3):414–414, 2002.

[26] Cyberspace Administration of China. Interim measures for the management of generative artificial intelligence services, 2023. Unofficial English translation; measures effective Aug 15, 2023.

[27] Department for Science, Innovation and Technology (UK). A pro-innovation approach to AI regulation, 2023. White Paper, March 2023.

[28] Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs (European Commission). Liability for defective products. `https://single-market-economy.ec.europa.eu/single-market/goods/free-movement-sectors/liability-defective-products_en`, Dec. 2024. Accessed: 2025-08-21.

[29] M. N. Duffourc and S. Gerke. The proposed eu directives for ai liability leave worrying gaps likely to impact medical ai. *NPJ Digital Medicine*, 6(1):77, 2023.

[30] S. Dutta, D. Wei, H. Yueksel, P.-Y. Chen, S. Liu, and K. R. Varshney. Is There a Trade-Off Between Fairness and Accuracy? A Perspective Using Mismatched Hypothesis Testing. Technical report, 2020.

[31] C. Dwork. Differential privacy. In *33rd International colloquium on automata, languages, and programming*, pages 1–12, Venice, Italy, 2006. Springer, Springer.

[32] L. E. Erdogan, N. Lee, S. Kim, S. Moon, H. Furuta, G. Anumanchipalli, K. Keutzer, and A. Gholami. Plan-and-act: Improving planning of agents for long-horizon tasks. *arXiv preprint arXiv:2503.09572*, 2025.

[33] ETH Zurich. Safe ai - laboratory for trustworthy ai systems, 2024. [Accessed: 2024-10-16].

[34] European Commission. Proposal for a directive on adapting non-contractual civil liability rules to artificial intelligence (ai liability directive), 2022.

[35] European Commission. Proposal for a directive on liability for defective products, 2022. Revises Council Directive 85/374/EEC.

[36] European Parliament and Council. Regulation (eu) 2016/679 (general data protection regulation). Official Journal of the European Union, 2016. OJ L 119, 4 May 2016, p. 1–88.

[37] European Parliament and Council. Regulation (eu) 2022/2065 on a single market for digital services (digital services act). Official Journal of the European Union, 2022. OJ L 277, 27 Oct 2022, p. 1–102.

[38] Federal Trade Commission. Aiming for truth, fairness, and equity in your company's use of AI, 2021. FTC Business Blog (Apr 19, 2021).

[39] Future of Life Institute. Eu artificial intelligence act – up-to-date developments and analyses of the eu ai act, 2025.

[40] G7. Hiroshima process international code of conduct for organizations developing advanced ai systems. G7 Digital Ministers' Hiroshima & Leaders' Process (Oct–Dec 2023), 2023.

[41] G7 Leaders. G7 leaders' statement on AI for prosperity. G7 Canada — Government of Canada, June 2025. Accessed 2025-08-21.

[42] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, A. Jones, S. Bowman, A. Chen, T. Conerly, N. DasSarma, D. Drain, N. Elhage, S. El-Showk, S. Fort, Z. Hatfield-Dodds, T. Henighan, D. Hernandez, T. Hume, J. Jacobson, S. Johnston, S. Kravec, C. Olsson, S. Ringer, E. Tran-Johnson, D. Amodei, T. Brown, N. Joseph, S. McCandlish, C. Olah, J. Kaplan, and J. Clark. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned, Nov. 2022. arXiv:2209.07858 [cs].

[43] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[44] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In T. Cohn, Y. He, and Y. Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online, Nov. 2020. Association for Computational Linguistics.

[45] Global Partnership on AI (GPAI). About the global partnership on artificial intelligence, 2020. International multi-stakeholder initiative hosted at the OECD.

[46] Government of Canada, Innovation, Science and Economic Development Canada. Artificial intelligence and data act (aida), bill c-27, digital charter implementation act, 2022, 2022.

[47] GPAI. Global partnership on artificial intelligence, 2024. [Accessed 18-11-2024].

[48] R. Hamon, H. Junklewitz, I. Sanchez, et al. Robustness and explainability of artificial intelligence. *Publications Office of the European Union*, 207:2020, 2020.

[49] M. U. Hassan, M. H. Rehmani, and J. Chen. Privacy preservation in blockchain based iot systems: Integration issues, prospects, challenges, and future research directions. *Future Generation Computer Systems*, 97:512–529, 2019.

[50] Humane Inc. Humane - technology built for people, 2024. [Accessed 2024-10-16].

[51] IEEE Standards Association. Ieee 7000™–2021: Standard model process for addressing ethical concerns during system design, 2021.

[52] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, D. Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

[53] Information Commissioner's Office (ICO). Ai and data protection risk toolkit. `https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/ai-and-data-protection-risk-toolkit/`, 2023. Guidance updated on 15 March 2023; under review due to Data (Use and Access) Act coming into law on 19 June 2025.

[54] International Telecommunication Union (ITU). Ai for good. `https://aiforgood.itu.int/`, 2024. [Accessed: 2024-09-25.

[55] ISO/IEC JTC 1/SC 42. Iso/iec 22989:2022 — artificial intelligence — concepts and terminology, 2022.

[56] ISO/IEC JTC 1/SC 42. Iso/iec 23053:2022 — framework for AI systems using machine learning, 2022.

[57] ISO/IEC JTC 1/SC 42. Iso/iec 23894:2023 — information technology — artificial intelligence — guidance on risk management, 2023.

[58] ISO/IEC JTC 1/SC 42. Iso/iec 42001:2023 — artificial intelligence management system (aims) — requirements, 2023.

[59] ISO/IEC JTC 1/SC 42. Iso/iec 5259–1:2024 — artificial intelligence — data quality for analytics and machine learning — part 1: Overview, terminology and examples, 2024. Foundational part of the ISO/IEC 5259 series.

[60] S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, et al. Language models (mostly) know what they know, 2022.

[61] B. Kuehnert, R. Kim, J. Forlizzi, and H. Heidari. The “who", “what", and “how" of responsible ai governance: A systematic review and meta-analysis of (actor, stage)-specific tools. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, page 2991–3005. ACM, June 2025.

[62] R. A. Leflar. The torts provisions of the restatement (second). *Columbia Law Review*, 72(2):267–278, 1972.

[63] B. Lepri, N. Oliver, and A. Pentland. Ethical machines: The human-centric use of artificial intelligence. *IScience*, 24(3), 2021.

[64] Y. Liu, G. Deng, Y. Li, K. Wang, Z. Wang, X. Wang, T. Zhang, Y. Liu, H. Wang, Y. Zheng, et al. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*, 2023.

[65] Y. Liu, J. Huang, Y. Li, D. Wang, and B. Xiao. Generative ai model privacy: A survey. *Artificial Intelligence Review*, 58(33), 2025.

[66] C. Lu, C. Lu, R. T. Lange, J. Foerster, J. Clune, and D. Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

[67] J. Lu, T. Holleis, Y. Zhang, B. Aumayer, F. Nan, F. Bai, S. Ma, S. Ma, M. Li, G. Yin, et al. Toolsandbox: A stateful, conversational, interactive evaluation benchmark for llm tool use capabilities, 2024.

[68] Q. Lu, L. Zhu, X. Xu, J. Whittle, D. Zowghi, and A. Jacquet. Responsible ai pattern catalogue: A collection of best practices for ai governance and engineering. *ACM Computing Surveys*, 56(7):1–35, 2024.

[69] D. Maclean. The nist risk management framework: Problems and recommendations. *Cyber Security: A Peer-Reviewed Journal*, 1(3):207–217, 2017.

[70] K. Madhavan, A. Yazdinejad, F. Zarrinkalam, and A. Dehghantanha. Quantifying security vulnerabilities: A metric-driven security analysis of gaps in current ai standards, 2025.

[71] Ministry of Economy, Trade and Industry (METI). AI guidelines for business (ver. 1.0), 2024. English version released 28 Mar 2024.

[72] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

[73] D. Monett and B. Grigorescu. Deconstructing the ai myth: Fallacies and harms of algorithmification. In *European Conference on e-Learning*, volume 23, pages 242–248, 2024.

[74] Montréal International. The global partnership on artificial intelligence officially launched. Montréal International, June 2020.

[75] National Institute of Standards and Technology. Nist privacy framework: A tool for improving privacy through enterprise risk management, version 1.0. Technical report, U.S. Department of Commerce, NIST, 2020.

[76] National Institute of Standards and Technology. Artificial intelligence risk management framework (ai rmf 1.0). Technical Report NIST AI 100-1, U.S. Department of Commerce, NIST, 2023.

[77] N. I. of Standards and T. (NIST). National institute of standards and technology, 2024. [Accessed 18-11-2024].

[78] OpenAI. Gpt-4 system card. Technical report, OpenAI, 2023. Accessed YYYY-MM-DD.

[79] Organisation for Economic Co-operation and Development (OECD). Recommendation of the council on artificial intelligence, 2019. Adopted 22–23 May 2019.

[80] Organisation for Economic Co-operation and Development (OECD. OECD AI Principles: Accountability (Principle 1.5), 2024. [Accessed 01-10-2024].

[81] C. C. Outeda. European education area and digital education action plan (2021–2027): One more step towards the europeanisation of education policy. In *E-Governance in the European Union: Strategies, Tools, and Implementation*, pages 187–206. Springer, 2024.

[82] C. C. Outeda. The eu's ai act: a framework for collaborative governance. *Internet of Things*, 27(1):101291, 2024.

[83] K. O'Sullivan, M. Markovic, J. Dymiter, B. Scheliga, C. Odo, and K. Wilde. Semi-automated data provenance tracking for transparent data production and linkage to enhance auditing and quality assurance in trusted research environments. *International Journal of Population Data Science*, 10(2):2464, 2025.

[84] Personal Data Protection Commission (Singapore). Model AI governance framework (second edition), 2020.

[85] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration. *arXiv preprint arXiv:1709.02012*, 2017.

[86] M. Pushkarna, A. Zaldivar, and O. Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1776–1826, 2022.

[87] P. T. Quarshie, A.-R. Abdulai, E. Duncan, K. B. Kc, R. Roth, A. Sneyd, and E. D. Fraser. Myth or reality? the digitalization of climate-smart agriculture (dcsa) practices in smallholding agriculture in the bono east region of ghana. *Climate Risk Management*, 42:100553, 2023.

[88] S. Rabanser, S. Günnemann, and Z. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. *Advances in Neural Information Processing Systems*, 32, 2019.

[89] I. D. Raji, E. M. Bender, A. Paullada, E. Denton, and A. Hanna. Ai and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*, 2021.

[90] S. Raza, A. Shaban-Nejad, E. Dolatabadi, and H. Mamiya. Exploring bias and prediction metrics to characterise the fairness of machine learning for equity-centered public health decision-making: A narrative review. *IEEE Access*, 2024.

[91] M. Sameki, S. Bird, and K. Walker. Interpretml: A toolkit for understanding machine learning models, 2020. [Accessed 30-09-2024].

[92] R. Schwartz, A. Vassilev, K. K. Greene, L. Perine, A. Burt, and P. Hall. Towards a standard for identifying and managing bias in artificial intelligence. Technical Report Special Publication 1270, National Institute of Standards and Technology (NIST), Mar. 2022. Published March 15, 2022; accessed August 21, 2025.

[93] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison. Hidden technical debt in machine learning systems. *Advances in neural information processing systems*, 28, 2015.

[94] T. Shi, J. He, Z. Wang, L. Wu, H. Li, W. Guo, and D. Song. Progent: Programmable privilege control for llm agents. *arXiv preprint arXiv:2504.11703*, 2025.

[95] P. Slattery, A. K. Saeri, E. A. Grundy, J. Graham, M. Noetel, R. Uuk, J. Dao, S. Pour, S. Casper, and N. Thompson. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence, 2024.

[96] I. I. S. M. Systems. ISO/IEC TS 27022:2021. Technical Report ISO/IEC TS 27022:2021, ISO - International Organization for Standardization, 3 2021.

[97] K. Tallam and E. Miller. Operationalizing camel: Strengthening llm defenses for enterprise deployment. *arXiv preprint arXiv:2505.22852*, 2025.

[98] The White House. Executive order 14110 — safe, secure, and trustworthy development and use of artificial intelligence. Federal Register, Oct 2023.

[99] TNO. Ai research - tno, 2024. [Accessed: 2024-10-16.

[100] H. Torfah, S. Junges, D. J. Fremont, and S. A. Seshia. Formal analysis of ai-based autonomy: from modeling to runtime assurance. In *International Conference on Runtime Verification*, pages 311–330. Springer, 2021.

[101] Treasury Board of Canada Secretariat. Directive on automated decision-making, 2021. Effective Apr 1, 2021; applies to Government of Canada institutions.

[102] UKRI Trustworthy Autonomous Systems (TAS) Hub. Ukri trustworthy autonomous systems (tas) hub. `https://tas.ac.uk/`, 2024. [Accessed: 2024-09-25.

[103] UNESCO. Recommendation on the ethics of artificial intelligence. UNESCO General Conference (41st), 23 Nov 2021, 2021.

[104] U.S. Government Accountability Office. Artificial intelligence: An accountability framework for federal agencies and other entities. Technical Report GAO-21-519SP, U.S. Government Accountability Office, June 2021. Published and publicly released on June 30, 2021; accessed on August 21, 2025.

[105] Y. Wang, R. Zheng, L. Ding, Q. Zhang, D. Lin, and D. Tao. Uncertainty aware learning for language model alignment. *arXiv preprint arXiv:2406.04854*, 2024.

[106] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[107] White House Office of Science and Technology Policy. Blueprint for an AI bill of rights: Making automated systems work for the american people, 2022.

[108] A. Winfield. Ethical standards in robotics and ai. *Nature Electronics*, 2(2):46–48, 2019.

[109] Q. Yang. Toward responsible ai: An overview of federated learning for user-centered privacy-preserving computing. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4):1–22, 2021.

[110] Q. Zhu. Game theory meets llm and agentic ai: Reimagining cybersecurity for the age of intelligent threats, 2025.

## Appendix

Table 5: AI governance controls: operationalization and auditable evidence.

| **Governance control** | **Operationalization (examples)** | **Auditable evidence (examples)** |
|---|---|---|
| Data governance & lineage | Datasheets; lineage tracking; access control; dataset update policy | Datasheets [43]; lineage & access logs; change records |
| Transparency artefacts | Model/System Cards; intended-use & limitations; risk disclosures | Model Cards [72]; System Cards [78, 16] |
| Operational logging & traceability | Prompt/output logs; tool-use traces; policy hit/override logging; retention schedule | Production-readiness checks [21]; RMF records [69]; debt risks [93] |
| Human oversight (HITL) | Review checkpoints; escalation paths; override controls; SLAs for sensitive actions | HITL guidelines [15]; oversight policy [63] |
| Policy enforcement & safety filters | Moderation pipelines; allow/deny lists; rate limits; periodic red-teaming | Filter eval reports (FN/FP); policy docs; guardrails [52] |
| Post-market monitoring | Drift detection; incident management; on-call runbooks; corrective actions | Monitoring dashboards; incident postmortems; RMF metrics [77] |
| Compliance documentation | Conformity dossier; risk register; residual-risk statements; audit trail | Design specs & logs; audit reports; AI Act alignment [82] |

Table 6: Comparison of selected AI governance and risk management frameworks, highlighting evaluation mechanisms and typical auditable evidence.

| **Framework** | **Type / scope** | **Core evaluation & assurance elements** | **Typical auditable evidence** |
|---|---|---|---|
| EU AI Act [82, 39] | Binding regulation (EU) | Risk classification; conformity assessment; QMS; lifecycle risk management; transparency/oversight; robustness/cybersecurity; post-market monitoring | Technical documentation; logs/retention; risk reports; incident reports |
| OECD AI Principles [80] | Non-binding principles (OECD) | Accountability; robustness/safety; transparency; fairness; risk–benefit assessment | Governance policies; impact assessments; transparency statements |
| NIST AI RMF 1.0 [69] | Voluntary risk framework (US) | Map–Measure–Manage–Govern; risk identification; metrics; controls; continuous monitoring | Risk register; measurement plan; monitoring KPIs; governance records |
| ISO/IEC 42001 (AIMS) [4] | Certifiable management system | Organization-level AI management requirements; process controls; continuous improvement | AIMS scope; procedures; internal audits; management reviews |
| GAO AI Accountability Framework [104] | Public-sector accountability | Governance; data; performance; monitoring | Accountability checklist; audit artifacts; change/control logs |
| UK ICO [11] & Alan Turing "Explaining AI" / AI Auditing | Regulator guidance (UK) | Explainability; documentation; audit processes; routes for redress | DPIAs/AIA-style docs; explanation records; audit plans |
| UK AI Cybersecurity Code of Practice [9] | Government code (UK) | Secure-by-design AI; threat modeling; controls; monitoring | Secure-by-design evidence; threat models; security test results |
| Canada AIDA [46] & Voluntary GenAI Code | Legislation (proposed) & code (Canada) | Risk/impact management; transparency; incident reporting | Impact assessments; notices; incident logs |
| WHO LMM Guidance for Health [10] | Sectoral guidance (Health) | Safety/efficacy; data governance; monitoring; clinical risk mgmt | Clinical validation; data lineage; post-deployment monitoring |
| Singapore AI Verify (Model Governance for GenAI) [8] | Model governance (GenAI) | Model governance controls; testing expectations; transparency | Governance checklists; testing summaries; release gates |
| ISO/IEC TR 24027 [2] (Bias/Data Quality) | Technical report (ISO) | Bias sources/metrics; data quality guidance | Bias assessments; data quality reports |
| NIST SP-1270 (Bias) [92] | Special publication (US) | Identify/manage bias; metrics; mitigation process | Bias measurement plans; subgroup analyses; mitigation evidence |
| MIT "AI Risk Repository" (crosswalk) [95] | Repository / synthesis | Cross-mapping of risks, controls, standards, practices | Control mappings; checklists; references to tests/benchmarks |

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly present the scope of our survey. They state four core contributions: (1) synthesizing workflow-level defenses for generative AI, (2) examining evaluation protocols and compliance criteria shaped by regulation, (3) analyzing liability and governance approaches, and (4) introducing the "AI Scientist" paradigm. Each of these is elaborated in dedicated sections of the paper (Sections 2–5) and revisited in the conclusion, ensuring consistency between claims and content

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We acknowledge that, as a survey, our work does not provide new empirical experiments but instead synthesizes and evaluates existing literature. While we aim for breadth across technical, regulatory, and governance perspectives, we note that some areas (e.g., rapidly evolving industry practices and region-specific regulations) could not be exhaustively covered. We also highlight that the "AI Scientist" paradigm remains aspirational and requires further empirical validation, which is beyond the scope of this paper.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper is a survey and does not introduce new theoretical results, theorems, or formal proofs. Instead, it synthesizes and critiques existing technical, governance, and regulatory literature. Where relevant, we reference prior work that provides formal definitions or proofs (e.g., fairness metrics, robustness evaluations), but no original theorems are contributed by this paper.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: This paper does not present original experimental results. Instead, it surveys and synthesizes existing literature on Responsible Generative AI, including technical safeguards, governance frameworks, and regulatory perspectives. Where we cite experimental work (e.g., on fairness metrics, robustness, or evaluation testbeds), reproducibility details are provided in the referenced papers, but no new experiments are introduced in this survey.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: This paper does not introduce new datasets, models, or code, as it is a survey. However, to support transparency and reproducibility of our synthesis, we will make the full set of references (bibliography) openly available, serving as the primary "data" of this survey. This allows readers to trace back all claims and analyses to the cited works. No additional code or experimental assets are required for the paper's contributions.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: This paper does not introduce new experiments or models. It surveys existing literature on Responsible Generative AI and therefore does not involve training/test setups or hyperparameter tuning.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: As a survey, the paper does not include original experiments. Where experimental findings from prior work are discussed, we reference the corresponding studies that provide their own statistical reporting.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: This paper does not present original experimental results, so no compute resources are reported. Instead, we synthesize findings from prior literature that may have used varied computational resources.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper is a literature survey and does not involve human or animal subjects, sensitive data collection, or potentially harmful interventions. All referenced works are properly cited, and the study fully adheres to the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both positive and negative societal impacts of Responsible Generative AI. On the positive side, we highlight how aligning technical safeguards, evaluation protocols, and governance frameworks can improve safety, transparency, and accountability in AI systems. On the negative side, we discuss risks such as bias, privacy leakage, misuse (e.g., disinformation, surveillance), and governance gaps if safeguards are not implemented. We also propose mitigation strategies, such as sandboxing, provenance tracking, and the "AI Scientist".

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not release new datasets or models. Instead, it surveys existing work and highlights safeguard mechanisms proposed in the literature, such as sandboxed execution, runtime monitoring, moderation pipelines, and regulatory oversight frameworks, but it does not itself release any high-risk assets.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All referenced works, datasets, models, and governance frameworks are cited to their original sources with appropriate bibliographic references (see References section). Since no external code or datasets are directly reused, there are no additional licensing concerns beyond proper scholarly citation.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This paper does not introduce new datasets, models, or code. The contribution is conceptual and survey-based, synthesizing existing literature and governance frameworks.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve human-subject research or crowdsourcing studies.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve experiments with human subjects and therefore does not require IRB approval.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: This survey does not use LLMs as part of its methodology or results. Any AI tools used were limited to standard writing/editing assistance and did not affect the scientific content or originality of the work.