Paper PP06

The SMART+ Framework for AI Systems

Laxmiraju Kandikatla, Maxis AI, Edison, NJ, United States

ABSTRACT

Artificial Intelligence (AI) systems are now integral to clinical research processes, such as automated adverse event detection in clinical trials, patient eligibility screening for protocol enrollment, and data quality validation. While these technologies enhance operational efficiency, they introduce new challenges regarding safety, accountability, and regulatory compliance. To address these concerns, we introduce the SMART+ Framework, a structured model built on the pillars of Safe, Monitored, Accountable, Reliable, Transparent, and augmented with Privacy, Data Governance, and Guardrails. SMART+ offers a practical, comprehensive approach to evaluating and governing AI systems within regulated environments. This framework aligns with GxP, ICH, and evolving regulatory guidance by integrating operational safeguards, oversight mechanisms, and strengthened privacy and governance controls. Through real-world scenarios, SMART+ demonstrates risk mitigation, trust-building, and compliance readiness. By enabling responsible AI adoption, protecting patient safety, and ensuring auditability, SMART+ provides a robust foundation for effective AI governance in clinical research.

1. INTRODUCTION

AI-driven tools and AI Systems agents hold significant promise in healthcare, from diagnostics to clinical trial optimization. However, these technologies also carry substantial risks if not carefully designed, validated, and monitored (Bouderhem, 2024; Ferrara et al., 2024; Khan et al., 2025). A high-profile example is the Epic Sepsis Model, a proprietary algorithm deployed in numerous hospitals, which failed to identify 67% of patients with sepsis while generating alerts for only 18% of admissions (Habib et al., 2021; Wong et al., 2021). This poor real-world performance far below clinician judgment highlights the potential consequences of relying on AI systems without rigorous oversight. Similarly, imaging AI models trained predominantly on light-skinned patients have demonstrated worse performance for lesions on darker skin tones, leading to underdiagnosis in non-white populations (Cross et al., 2024; Rezk et al., 2022). These cases illustrate how AI systems that perform well in controlled environments can fail in practice, potentially exacerbating healthcare disparities and compromising patient safety.

2. METHODOLOGICAL CONSIDERATIONS

Several established trustworthy AI frameworks provide complementary guidance on the development and governance of AI systems. The NIST AI Risk Management Framework offers a comprehensive, risk-based approach across the AI lifecycle, emphasizing characteristics such as validity, reliability, safety, security, transparency, and accountability (NIST, 2023). The OECD AI Principles promote innovative and responsible AI aligned with human rights, democratic values, and inclusive growth (OECD, 2019). The U.S. Government Accountability Office (GAO) AI Accountability Framework outlines practical governance, data management, performance evaluation, and continuous monitoring practices, particularly for public-sector AI systems (GAO, 2021). Similarly, the EU Ethics Guidelines for Trustworthy AI define a set of requirements—such as human agency, technical robustness, privacy, and societal well-being—that AI systems should meet to be considered trustworthy (EU, 2019). While these frameworks collectively advance the state of trustworthy AI, they do not fully address the domain-specific challenges associated with clinical research and healthcare deployment, including patient safety, regulatory compliance, and human oversight in high-risk decision-making contexts (WHO, 2021).

This SMART+ Framework introduces a streamlined taxonomy tailored for healthcare AI. It integrates the core principles of Safe, Monitored, Accountable, Reliable, and Transparent (SMART), augmented with Privacy & Security, Data Governance, Fairness & Bias, and Guardrails. By doing so, SMART+ provides actionable guidance to ensure that AI systems in clinical research operate ethically, reliably, and safely. We aim to provide a practical framework for achieving trustworthiness in AI systems by ensuring that all essential parameters (SMART+) are thoroughly addressed before the system is released for real-world use. Beyond deployment, our framework emphasizes the importance of continuous monitoring to maintain performance, safety, and ethical integrity over time. This holistic approach enables stakeholders to apply well-defined principles for evaluating AI systems and demonstrating their trustworthiness throughout the entire lifecycle.

The framework recognizes that trust in AI cannot be achieved through isolated checks; rather, it must be systematically embedded across the entire AI lifecycle. Accordingly, we consider all major phases—objective setting, requirements and specifications, design and development, verification and validation, deployment, and operation and maintenance—as interdependent stages where trustworthiness must be intentionally cultivated. Within each phase, we incorporate the SMART+ principles as a unifying structure to ensure that expectations are clearly defined, measurable, actionable, and aligned with ethical and technical quality standards.

A central component of the framework is the integration of risk assessment to account for system complexity, context of use, and potential harm. The outcome of this assessment informs the proportional application of SMART+ principles: higher-risk AI systems warrant augmented controls, enhanced validation rigor, and more robust oversight, whereas lower-risk systems require correspondingly lighter governance. This risk-stratified approach ensures that resources are directed where they are most needed while maintaining adherence to principles of responsible innovation. By combining lifecycle orientation, principle-based design, and risk-proportionate governance, the proposed framework offers a holistic and traceable method for evaluating trustworthiness. Moreover, it supports continuous monitoring after deployment to ensure that AI systems remain safe, reliable, and aligned with stakeholder expectations throughout their operational lifespan

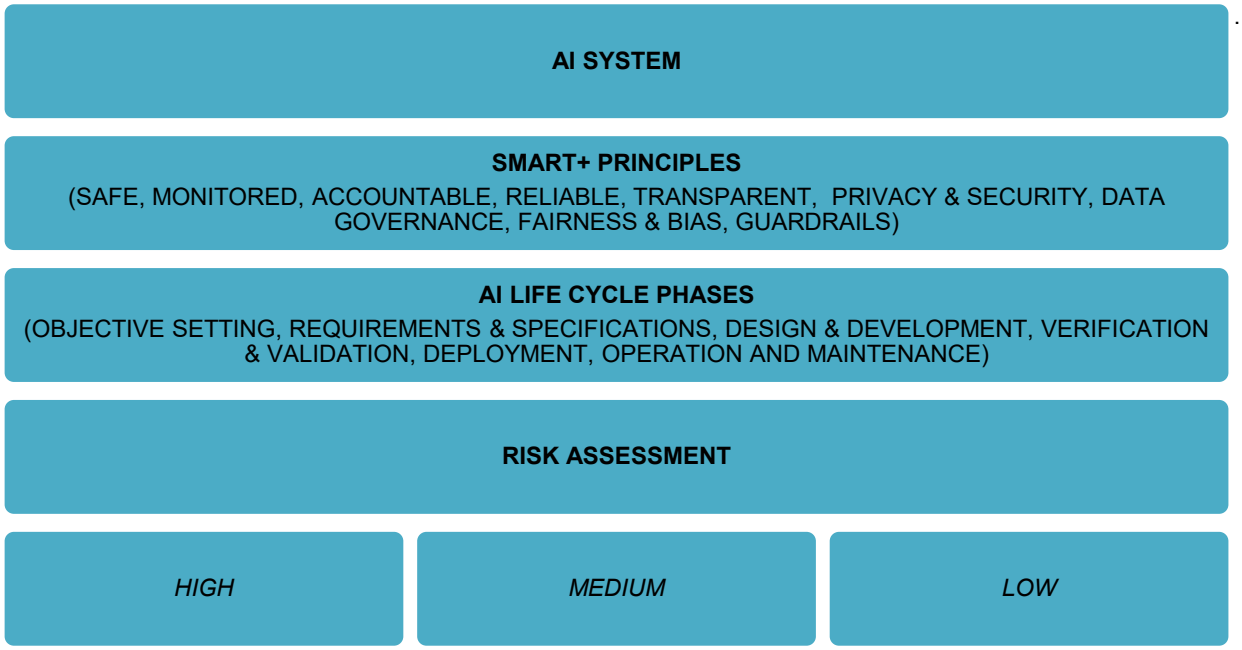

Figure 1: SMART + FRAMEWORK

3. SMART + PRINCIPLES

The SMART+ Framework is a structured model designed to ensure the development, deployment, and monitoring of trustworthy AI systems in regulated and high-stakes domains, such as healthcare, clinical research, and pharmaceuticals. It extends traditional AI governance approaches by integrating multiple dimensions of system quality, accountability, and ethical compliance, each representing a critical aspect of AI trustworthiness (Figure 2).

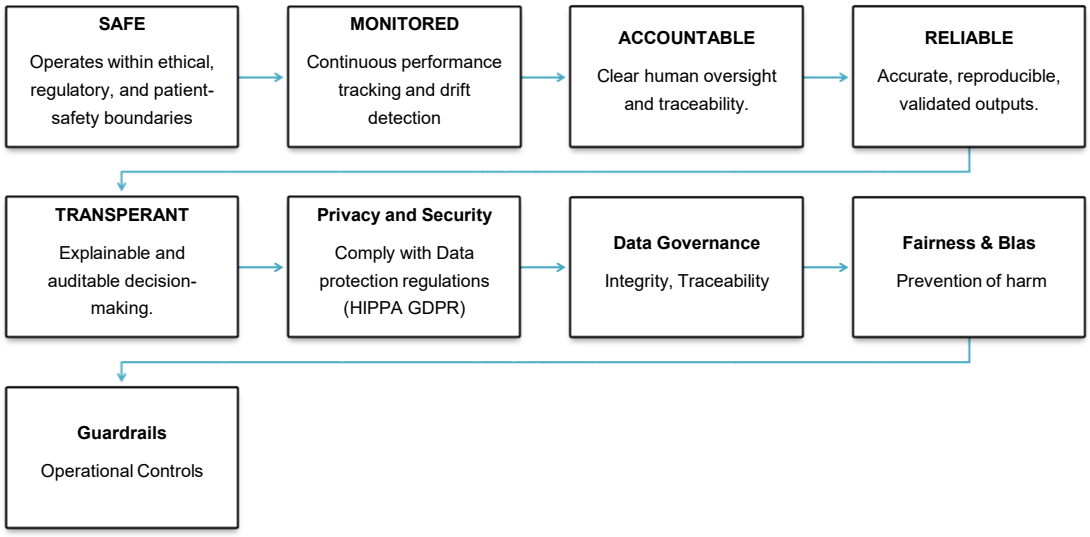

Figure 2. SMART+ Principles

By explicitly defining standards and checkpoints under each pillar, SMART+ provides a comprehensive mechanism to identify, mitigate, and monitor risks across the AI lifecycle, aligning with global guidelines. Its structured approach allows organizations to systematically assess AI systems, ensuring operational safety, data integrity, ethical compliance, and regulatory readiness. In other words, SMART+ can be systematically embedded into the AI lifecycle. It translates ISO 42001's principles into a set of operational dimensions that shape both practice and assurance in contemporary AI management (Figure 3).

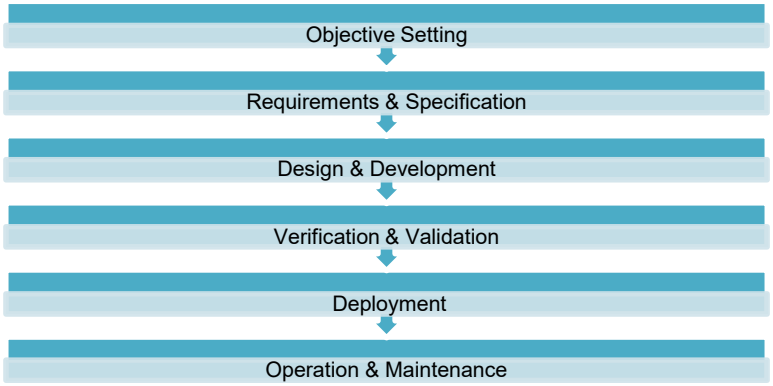

Figure 3. AI System Life Cycle Management

4. PRACTICAL APPROACH

4.1 SAFETY

The Safety pillar represents the foundational ethical and technical commitment to ensure that AI systems operate without causing harm to individuals, society, or the environment. Safety in AI encompasses the prevention of unintended consequences, robustness against adversarial manipulation, and reliability under varying operational conditions (Ferrara et al., 2024). As defined by the EU Ethics Guidelines for Trustworthy AI (EU, 2019) and the NIST AI Risk Management Framework (2023), safety entails both technical robustness (resilience, accuracy, and reliability) and ethical soundness (avoidance of harm, bias, or misuse). In high-stakes domains such Clinical Research Industry, safety aligns with the "do no harm" principle and the ISO 31000 and ISO 23894 approach to risk management (ISO, 2018 & ISO, 2023) where potential failure mode is anticipated, quantified, and controlled. Regarding its integration into the AI lifecycle, Safety must be embedded as a continuous assurance activity throughout all phases:

- Objective Setting, through the definition of the AI system's purpose, ethical boundaries, and potential risk domains, as well as the identification of critical safety use cases where harm may occur (e.g., patient misdiagnosis, erroneous recommendations).
- Requirements & Specification, through the establishment of explicit safety requirements and risk tolerance thresholds, as well as the specifications of failsafe mechanisms, fallback procedures, and performance validation criteria.
- Design & Development, through the incorporation of safety by design principles, redundancy mechanisms, and guardrail configurations during model training and testing, as well as the implementation of bias detection, adversarial robustness testing, and safety validation checkpoints.
- Verification & Validation, through stress testing, scenario simulation, and independent validation to confirm safe performance under edge cases, as well as the verification of conformance to regulatory and ethical safety standards (e.g., ISPE GAMP 5 (ISPE, 2022), ISO/IEC 5338 (ISO, 2023a)).
- Deployment, through the implementation of human oversight mechanisms for critical decision points, as well as the documentation of incident response plans and establishment of rollback procedures; and
- Operation & Maintenance, through the continuous monitoring of AI behavior, performance drifts, and incident reports, as well as periodic safety audits and integration of lessons learned into retraining cycles.

4.2 MONITORING

The Monitored pillar ensures that AI systems are continuously observed for performance consistency, ethical compliance, and operational safety throughout their lifecycle. Monitoring establishes a feedback mechanism for detecting drift, bias, security anomalies, or system degradation in real time (Corrêa et al., 2023; Papagiannidis et al., 2025; Radanliev, 2025). In the context of AI, monitoring extends beyond technical performance to include ethical performance—tracking unintended consequences, data shifts, and user feedback that may affect fairness, safety, and accountability. Therefore, monitoring serves as the dynamic quality control loop that sustains trustworthy AI during both development and operational use. Monitoring outlines AI Safety Events Log, where each row records the system identity, event details, severity level, root cause, remediation, and audit notes. Such structured logging enables trend analysis and regulatory reporting. By continuously testing AI (e.g., via shadow mode or simulated data)

and incorporating user feedback, organizations catch issues early. This aligns with EU/OECD calls for feedback mechanisms and human oversight loops.

Regarding its integration into the AI lifecycle, Monitoring is applied to multiple phases, ensuring proactive detection and mitigation of deviations from expected behavior:

- Objective Setting, through the definition of measurable key performance indicators (KPIs), alert thresholds, and monitoring objectives aligned with risk categories (e.g., high-risk AI systems per EU AI Act Article 9).
- Requirements & Specification, through the inclusion of monitoring and audit requirements (e.g., system logging, data drift detection, explainability audits) in the system specifications.
- Design & Development, through the integration of built-in monitoring functions such as model monitoring dashboards, bias detection pipelines, and safety triggers during model evaluation.
- Verification & Validation, through the validation of monitoring functions by conducting stress and robustness tests under controlled failure conditions (simulated anomalies or out-of-range inputs).
- Deployment, through the introduction of monitoring dashboards, alert mechanisms, and automated logging systems, as well as the definition of escalation paths and responsible personnel for anomaly management; and
- Operation & Maintenance, through the performance of continuous monitoring for model drift, accuracy degradation, and ethical compliance, as well as periodic system audits, retraining reviews, and governance board oversight meetings to assess ongoing trustworthiness.

4.3 ACCOUNTABILITY

The Accountable pillar embodies the principle that responsibility for AI outcomes ultimately lies with humans and organizations, not the autonomous system itself. Accountability ensures that every AI decision, model modification, and system action is traceable, auditable, and attributable to defined human roles (Schuett et al., 2025; Solove & Matsumi, 2024). As emphasized by the EU AI Act (Articles 14 & 15), ISO/IEC 42001, and the NIST AI Risk Management Framework, accountability requires clear governance structures, transparent documentation, and mechanisms for redress in case of failure or harm. In regulated environments, accountability aligns with GxP (good practice) principles where each activity—design, testing, deployment, or monitoring—must be verified, approved, and justified by a qualified authority. Thus, accountability operationalizes trust through governance, transparency, and continuous oversight across the AI lifecycle. Regarding its integration into the AI lifecycle, Accountability should be established and reinforced at each phase through well-defined roles, decision checkpoints, and documentation controls.

- Objective Setting, through the definition of governance ownership, including AI sponsors, validation leads, and data stewards, as well as the establishment of a RACI (Responsible–Accountable–Consulted–Informed) matrix to delineate accountability boundaries.
- Requirements & Specification, through the documentation of data provenance, intended use, as well as the sign-off on ethical compliance and system purpose statements.
- Design & Development, through the maintenance of detailed design history files and version-controlled artifacts that capture model changes, rationale, and developer sign-offs.
- Verification & Validation, through the enforcement of independent quality review, traceability matrices, and deviation logs, whereby validation reports must be authorized by accountable personnel (Quality Assurance or Validation Lead).
- Deployment, through the approval for release that should follow a structured Change Control and Release Authorization process under the Quality Management System (QMS), and include rollback and escalation procedures; and
- Operation & Maintenance, through the maintenance of post-deployment audit trails, incident management logs, and retraining documentation, as well as periodic reviews of accountability structures as part of governance meetings or AI Ethics Board reviews.

4.4 RELIABILITY

The Reliable pillar ensures that an AI system performs consistently, accurately, and predictably across different environments, data sources, and use conditions. Reliability in AI reflects the system's ability to maintain intended functionality over time while minimizing unexpected outcomes or failures (Floridi et al., 2018). According to the NIST AI Risk Management Framework, reliability is a core attribute of trustworthy AI, emphasizing robustness, reproducibility, and resistance to drift or degradation. Similarly, ISO/IEC 42001 highlight reliability as a validation outcome, where systems are proven fit for intended use and remain stable throughout their operational lifecycle. In essence, a reliable AI system demonstrates technical robustness, performance stability, and repeatability under real-world and stress conditions, ensuring consistent value delivery without compromising safety or ethics. Overall, Reliability thus acts as the stability assurance bridge between AI model development and its safe, sustained operation. Regarding its integration into the AI lifecycle, Reliability applies across all phases, with particular emphasis on Design & Development, Verification & Validation, and Operation & Maintenance.

- Objective Setting, through the definition of performance expectations and reliability targets such as model accuracy, latency, reproducibility, and robustness criteria, as well as the alignment of reliability metrics with business and regulatory goals (e.g., EU AI Act Article 15).

- Requirements & Specification, through the specification of measurable performance criteria, validation conditions, and test environments in system and data requirements documents.
- Design & Development, through the incorporation of reliability engineering principles by building multiple models, testing alternative algorithms, and ensuring reproducibility across datasets and environments, as well as the conduct of robustness tests to evaluate performance under edge cases, missing data, and adversarial conditions;
- Verification & Validation, through the performance of formal validation and stress testing to confirm consistent results, as well as the use of statistical verification, confidence intervals, and benchmark comparisons to verify reproducibility, and the implementation of cross-validation and independent testing across temporal or demographic datasets.
- Deployment, through the validation of model reliability in real-world conditions via shadow deployment, pilot trials, or A/B testing before full rollout; and
- Operation & Maintenance, through the continuous monitoring of key performance metrics and data drift to identify reliability degradation, as well as the implementation of version control and retraining protocols to maintain performance over time.

.

4.5 TRANSPARENCY

The Transparency pillar ensures that the design, operation, and decision-making processes of an AI system are understandable, explainable, and traceable to both technical and non-technical stakeholders. Transparency in AI promotes visibility into how models' function, how data is processed, and how outcomes are derived, enabling accountability and trust across the system's lifecycle (Radanliev, 2025; Schmidt et al., 2020). According to the OECD AI Principles and the EU AI Act (Article 13), AI systems must be sufficiently transparent to enable users to interpret the system's output appropriately. The NIST AI Risk Management Framework similarly identifies transparency as a key property of trustworthy AI, emphasizing interpretability, traceability, and documentation. Likewise, ISO/IEC 42001 promotes transparency through structured documentation, model cards, and explainability records that support auditability and informed oversight. In essence, transparency serves as the connective tissue between AI developers, users, and regulators, ensuring that system logic, performance, and limitations are visible and communicated effectively. It forms the foundation for human oversight, risk management, and ethical assurance in AI-driven environments. Transparency must be embedded throughout the AI lifecycle to ensure consistent visibility and interpretability at each phase:

- Objective Setting, through the definition of transparency goals, such as explainability, traceability, and disclosure obligations, as well as the identification of stakeholders who require visibility (e.g., developers, regulators, end users) and the specification of communication needs.
- Requirements & Specification, through the inclusion of transparency requirements in functional and non-functional specifications (e.g., the need for model documentation, explainable AI (XAI) features, or audit trails), as well as the definition of how traceability will be maintained from data to decisions.
- Design & Development, through the incorporation of interpretable model architectures and documentation of the rationale for algorithmic choices, feature selection, and preprocessing steps, as well as the maintenance of data lineage and metadata to support reproducibility and accountability.
- Verification & Validation, through the conduct of explainability assessments to ensure model decisions can be justified, as well as the validation of documentation completeness, traceability matrices, and compliance with disclosure standards.
- Deployment, through the assurance that end-user interfaces present outputs in a comprehensible and non-deceptive manner, as well as the provision of users with information about the AI system's capabilities, limitations, and data usage; and
- Operation & Maintenance, through the continuous updating of transparency documentation as models evolve, as well as the maintenance of logs and audit records for retraining events, performance shifts, and decision rationales, and provision of traceability for any post-deployment modifications.

4.6 PRIVACY AND SECURITY

Privacy and Security form the foundational elements of trustworthy and compliant AI systems, ensuring that personal, sensitive, or proprietary data are collected, processed, and stored responsibly throughout the AI lifecycle. Privacy focuses on safeguarding individual data rights and ensuring compliance with data protection laws such as GDPR and HIPAA (Fabiano, 2025; Radanliev, 2025), while security emphasizes the protection of data and systems from unauthorized access, manipulation, or cyber threats (Alhitmi et al., 2024; Humphreys et al., 2024; Jada & Mayayise, 2024; Villegas-Ch & García-Ortiz, 2023). In the context of AI, privacy and security are critical not only for maintaining regulatory compliance but also for fostering public trust in algorithmic decision-making. The emergence of agentic AI systems further heightens these concerns, as autonomous decision loops can amplify data exposure risks or create new attack vectors if not properly governed. Privacy and security considerations must be embedded across every phase of the AI lifecycle:

- Objective Setting, through the definition of data protection objectives, access control policies, and regulatory compliance requirements (e.g., GDPR, ISO/IEC 27001, NIST AI RMF), as well as the conduct of a Data Protection Impact Assessment (DPIA), where applicable.

- Requirements & Specification, through the establishment of security- and privacy-by-design principles, encryption standards, and data minimization requirements, as well as the specification of auditability, traceability, and access management features.
- Design & Development, through the implementation of data anonymization, differential privacy, federated learning, and secure model development practices, as well as the integration of threat modeling techniques to identify potential vulnerabilities and misuse scenarios.
- Verification & Validation, through the performance of security testing (e.g., penetration testing, vulnerability scanning) and privacy compliance validation, as well as the validation of adherence to data governance controls and the assessment of resilience to data leakage or model inversion attacks.
- Deployment, through the enforcement of strong authentication, role-based access control, and network security measures, as well as the conduct of real-time monitoring for intrusion detection and anomaly identification; and
- Operation & Maintenance, through the continuous monitoring of privacy breaches and security vulnerabilities, as well as the implementation of incident response plans, retraining protocols, and regular patch management to maintain integrity and trustworthiness.

4.7 DATA GOVERNANCE

Data governance encompasses the policies, processes, and accountability frameworks that ensure data is managed ethically, securely, and in compliance with regulatory and organizational requirements throughout the AI lifecycle. It ensures that data used to train, validate, and operate AI systems is accurate, consistent, traceable, and fit for its intended purpose (Corrêa et al., 2023; Law & McCall, 2024; Tallberg et al., 2023). In the context of AI systems, effective data governance underpins trustworthiness, supporting fairness, privacy, and accountability. It mitigates risks such as data bias, data drift, and data misuse—all of which can directly affect model performance and regulatory compliance. The OECD AI Principles and the EU AI Act emphasize strong governance of data and data processes as essential to developing reliable and human-centric AI. Data governance must be woven across each phase of the AI lifecycle, ensuring that every data-related activity is controlled and auditable:

- Objective Setting, through the definition of governance objectives aligned with organizational policies, ethical AI guidelines, and applicable regulations (e.g., GDPR, ISO/IEC 38505), as well as the establishment of data ownership, accountability roles, and governance committees.
- Requirements & Specification, through the specification of data governance requirements such as metadata documentation, lineage tracking, consent management, and data quality standards, as well as the development of a data governance plan detailing lifecycle control mechanism.
- Design & Development, through the implementation of procedures for data acquisition, labeling, and curation aligned with governance policies, as well as the assurance of data traceability and version control, and the application of checks for data representativeness and bias mitigation before model training.
- Verification & Validation, through the validation of data integrity, completeness, and provenance, as well as the verification of compliance with data retention and anonymization requirements, and the performance of governance audits to ensure adherence to approved data sources and usage policies.
- Deployment, through the establishment of monitoring pipelines to track ongoing data inputs, ensuring that operational data adheres to governance standards and that no unauthorized data sources are introduced; and
- Operation & Maintenance, through the continuous monitoring for data drift, data quality degradation, and unauthorized data reuse, as well as the conduct of periodic data governance reviews and the maintenance of auditable logs for traceability and accountability.

4.8 FAIRNESS AND BIAS

Fairness and Bias reflect the ethical and technical imperative to ensure that AI systems provide equitable, non-discriminatory, and context-appropriate outcomes for all users and populations. Fairness in AI goes beyond simple accuracy parity—it encompasses the identification, mitigation, and continuous monitoring of biases that may arise from datasets, model architectures, workflows, or human–AI interactions (Barredo Arrieta et al., 2020; Mehrabi et al., 2021; NIST, 2023). As emphasized by the EU Ethics Guidelines for Trustworthy AI (EU, 2019) and the OECD AI Principles (OECD, 2021), fairness requires the prevention of both allocative harm (unequal distribution of opportunities or resources) and representational harm (reinforcement of stereotypes or exclusion of minority groups). Bias can be introduced through multiple vectors—historical biases embedded in training data, underrepresentation of subgroups, model design decisions, feedback loops, and operational context shifts. Therefore, fairness must be treated as a multi-dimensional risk factor requiring systemic controls, ongoing measurement, and periodic recalibration. The NIST AI RMF (2023) highlights the necessity of socio-technical evaluation, calling for domain-specific fairness metrics, transparent reporting, and context-appropriate definitions of equity. Regarding its integration into the AI lifecycle, Fairness and Bias mitigation must be embedded continuously across every phase

- Objective Setting, through the definition of fairness goals relevant to the intended use case, identification of sensitive attributes (e.g., age, gender, ethnicity), and assessment of potential differential impacts on subpopulations. This step also includes early ethical impact assessment and stakeholder consultation to identify equity-critical scenarios.

- Requirements & Specification, through the establishment of explicit fairness requirements, selection of appropriate fairness metrics (e.g., equalized odds, demographic parity, predictive parity), and documentation of unacceptable bias thresholds. This phase also includes requirements for inclusive data source and representativeness.
- Design & Development, through the integration of bias detection techniques, dataset balancing strategies, augmentation methods, and model-agnostic fairness interventions (e.g., reweighting, adversarial debiasing). Transparent model design choices, interpretability mechanisms, and subgroup performance reporting are essential outputs.
- Verification & Validation, through systematic fairness testing—including cross-group error analysis, sensitivity analysis, and scenario-based evaluations to capture hidden or emergent biases. Independent review, testing against baseline inequity thresholds, and validation against regulatory/ethical guidelines (e.g., GDPR, FDA Good Machine Learning Practice principles) are required to ensure fairness conformance.
- Deployment, through the implementation of human oversight checkpoints specifically designed to detect potential biased outputs in real workflows, especially in high-impact decision pathways. Deployment must also ensure traceability of fairness-related decisions and documentation of mitigation strategies, exceptions, or manual overrides; and
- Operation & Maintenance, through continuous fairness monitoring, automated alerts for subgroup performance drift, and periodic revalidation as demographic patterns or usage contexts evolve. Post-deployment audits, stakeholder feedback loops, and real-world impact assessments are essential for sustained fairness performance. Update cycles must incorporate fairness metrics alongside accuracy and safety metrics to avoid long-term systemic inequity.

4.9 GUARDRAILS

Guardrails in AI systems refer to the technical, procedural, and ethical boundaries that constrain AI behavior to ensure it operates within safe, lawful, and intended limits. They serve as proactive mechanisms for risk prevention and mitigation, ensuring that AI systems remain aligned with human values, regulatory requirements, and organizational objectives (Berkey, 2024; Hakim et al., 2025; Pandya et al., 2025). Guardrails function both as design-time safeguards (e.g., data constraints, validation checks) and runtime controls (e.g., human-in-the-loop oversight, anomaly detection). By embedding guardrails, organizations establish fail-safe mechanisms that uphold trust, accountability, and ethical alignment, resonating with frameworks like the EU Ethics Guidelines for Trustworthy AI and the NIST AI Risk Management Framework, which emphasize “human oversight” and “risk management through governance controls.” Guardrails must be integrated as cross-cutting controls across the AI lifecycle—from concept to post-deployment—to maintain continuous oversight and system resilience:

- Objective Setting, through the definition of the ethical, safety, and compliance boundaries for the AI system, as well as the identification of unacceptable outcomes and the establishment of guardrail objectives consistent with organizational policies and legal frameworks (e.g., GDPR, FDA, ISO/IEC 23894);
- Requirements & Specification, through the translation of ethical and regulatory expectations into functional and non-functional requirements, as well as the specification of acceptable performance thresholds, risk tolerance levels, and human intervention points.
- Design & Development, through the implementation of technical guardrails such as input validation, access control, model explainability, and safe response mechanisms, as well as the integration of policy-based constraints into training pipelines (e.g., excluding harmful or biased data) and the incorporation of oversight protocols like Human-in-Command (HIC), Human-in-the-Loop (HITL), and Human-on-the-Loop (HOTL) to ensure supervised autonomy;
- Verification & Validation, through the validation that guardrails perform as intended under normal and stress conditions, as well as the conduct of adversarial testing, scenario simulation, and edge-case evaluations to confirm robustness, and the documentation of results through guardrail verification reports.
- Deployment, through the implementation of runtime guardrails such as real-time monitoring, output filters, and fail-safe triggers that prevent the propagation of unsafe or noncompliant outputs, as well as the establishment of escalation mechanisms for human review of flagged events; and
- Operation & Maintenance, through the continuous monitoring of AI system behavior and adjustment of guardrails based on performance logs, incident trends, and emerging risks, as well as periodic review of thresholds and oversight mechanisms to align with updated policies, standards, or risk assessments.

5. CONCLUSION

The responsible development and deployment of AI systems in regulated domains such as healthcare requires a structured, lifecycle-oriented approach that embeds governance principles into every phase of system evolution. By operationalizing trustworthy AI pillars into actionable lifecycle, organizations can establish a measurable framework that moves beyond theoretical ethics commitments to demonstrable compliance evidence. In other words, SMART+ principles mapped across the lifecycle mandates early clarity of risk boundaries, safety-by-design modelling choices, independent validation before release, and ongoing surveillance for unintended outcomes. Collectively, these pillars form a governance scaffold that aligns with leading frameworks—including NIST AI RMF, ISO/IEC 42001, ISO/IEC 5338, EU AI Act, and ISPE GAMP 5 by translating abstract ethical expectations into practical validation controls.

Importantly, they establish a dynamic feedback ecosystem linking design, deployment, and post-market learning, acknowledging that AI is not a static artifact but an adaptive system requiring ongoing oversight.

Additionally, AI systems are categorized into low, medium, and high-risk levels using the model influence and decision-consequence model. Once the risk level is determined, the corresponding governance expectations are mapped to the SMART+ Framework. High-risk systems require the full application of all SMART+ items with the highest level of stringency, ensuring comprehensive controls across data, model, monitoring, and technical safeguards. Medium-risk systems mandate a moderate implementation of SMART+ controls, focusing on essential requirements while allowing context-based flexibility for advanced items. Low-risk systems require only a minimal or optional application of SMART+ elements, where the primary obligation is to conduct a basic risk assessment and apply lightweight controls as appropriate.

This tiered approach ensures that oversight scales proportionally with the potential impact and avoids unnecessary burden on low-impact systems while maintaining strong safeguards where needed most. Going forward, the integration of these pillars into quality management systems and AI management systems represents a significant step toward evidence-based AI governance.

REFERENCES

Alhitmi, H.K., Mardiah, A., Al-Sulaiti, K.I., & Abbas, J. (2024). Data security and privacy concerns of AI-driven marketing in the context of economics and business field: An exploration into possible solutions. *Cogent Business & Management*, 11(1), 2393743. https://doi.org/10.1080/23311975.2024.2393743

Berkey, F. (2024). Guardrails: Guiding human decisions in the age of AI. *Family Medicine*, 56(7), 462–463. https://doi.org/10.22454/FamMed.2024.383539

Bouderhem, R. (2024). Shaping the future of AI in healthcare through ethics and governance. *Humanities and Social Sciences Communications*, 11(416). https://doi.org/10.1057/s41599-024-02894-w

Corrêa, N.K., Galvão, Santos, J.W., et al. (2023). Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns*, 4(10), 100857. https://doi.org/10.1016/j.patter.2023.100857

Cross, J.L., Choma M.A., & Onofrey, J.A. (2024). Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health*, 3(11), e0000651. https://doi.org/10.1371/journal.pdig.0000651

EU. (2019). Ethics Guidelines for Trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

EU. (2024). EU Artificial Intelligence Act. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689

Fabiano, N. (2025). AI Act and Large Language Models (LLMs): When critical issues and privacy impact require human and ethical oversight. *arXiv*. https://doi.org/10.48550/arXiv.2404.00600

Ferrara, M., Bertozzi, G., Di Fazio, N., et al. (2024). Risk management and patient safety in the artificial intelligence era: A systematic review. *Healthcare*, 12(5). https://doi.org/10.3390/healthcare12050549

Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). *AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations*. *Minds and Machines*, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

GAO AI Accountability Framework. (2021). https://www.gao.gov/products/gao-21-519sp

Habib, A.R., Lin, A.A., & Grant, R.W. (2021). The epic sepsis model falls short: The importance of external validation. *JAMA Internal Medicine*, 181(8), 1040–1041. https://doi.org/10.1001/jamainternmed.2021.3333

Hakim, J.B., Painter, J.L., Ramcharran, D., et al. (2025). The need for guardrails with large language models in pharmacovigilance and other medical safety critical settings. *Scientific Reports*, 15, 27886. https://doi.org/10.1038/s41598-025-09138-0

Humphreys, D., Koay, A., Desmond, D., & Mealy, E. (2024). AI hype as a cyber security risk: The moral responsibility of implementing generative AI in business. *AI and Ethics*, 4(3), 791–804. https://doi.org/10.1007/s43681-024-00443-4

ISO. (2019). ISO 14971. https://www.iso.org/standard/72704.html

ISO. (2023a). ISO/IEC 42001. https://www.iso.org/standard/42001

ISPE. (2022). GAMP 5. https://guidance-docs.ispe.org/doi/book/10.1002/9781946964854

Jada, I., & Mayayise, T.O. (2024). The impact of artificial intelligence on organisational cyber security: An outcome of a systematic literature review. *Data and Informational Management*, 8(2), 100063. https://doi.org/10.1016/j.dim.2023.100063.

Khan, M.M., Shah, N., Shaikh, N., et al. (2025). Towards secure and trusted AI in healthcare: A systematic review of emerging innovations and ethical challenges. *International Journal of Medical Informatics*, 195, 105780. https://doi.org/10.1016/j.ijmedinf.2024.105780

Law, T., & McCall, L. (2024). Artificial intelligence policymaking: An agenda for sociological research. *Socius*, 10. https://doi.org/10.1177/23780231241261596

NIST AI Risk Management Framework. (2023). https://www.nist.gov/itl/ai-risk-management-framework

OECD AI Principles. (2019/2024). https://oecd.ai/en/ai-principles

Pandya, R., Bland, M., Nguyen, D.P., et al. (2025). From refusal to recovery: A control-theoretic approach to Generative AI guardrails. *arXiv*. https://doi.org/10.48550/arXiv.2510.13727

Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *Journal of Strategic Informational Systems*, 34(2), 101885. https://doi.org/10.1016/j.jsis.2024.101885

Radanliev, P. (2025). AI ethics: Integrating transparency, fairness, and privacy in AI development. *Applied Artificial Intelligence*, 39(1), 2463722. https://doi.org/10.1080/08839514.2025.2463722

Rezk, E., Eltorki, M., & El-Dakhakhni, W. (2022). Improving skin color diversity in cancer detection: Deep learning approach. *JMIR Dermatology*, 5(3), e39143. https://doi.org/10.2196/39143

Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, 29(4), 260–278. https://doi.org/10.1080/12460125.2020.1819094

Schuett, J., Reuel, A.K., & Carlier, A. (2025). How to design an AI ethics board. *AI and Ethics*, 5(2), 863–881. https://doi.org/10.1007/s43681-023-00409-y

Solove, D.J., & Matsumi, H. (2024). *AI, algorithms, and awful humans*. Fordham Law Review, 92(5), 1923–1940. https://ir.lawnet.fordham.edu/flr/vol92/iss5/8

Tallberg, J., Erman, E., Furendal, M., et al. (2023). The global governance of artificial intelligence: Next steps for empirical and normative research. *International Studies Review*, 25(3), viad040, https://doi.org/10.1093/isr/viad040

UNESCO. (2021). Recommendation on the Ethics of AI. https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence

Villegas-Ch., W., & García-Ortiz, J. (2023). Toward a comprehensive framework for ensuring security and privacy in artificial intelligence. *Electronics*, 12(18), 3786. https://doi.org/10.3390/electronics12183786

WHO. (2021). Ethics and Governance of AI for Health. https://www.who.int/publications/i/item/9789240029200

Wong, A., Otles, E., Donnelly, J.P., et al. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*, 181(8), 1065–1070. https://doi.org/10.1001/jamainternmed.2021.2626

CONTACT INFORMATION

Author Name: Laxmiraju Kandikatla
Company: 510 THORNALL STREET 180, Edison, NJ 08837
Address: Edison, NJ, United States
Work Phone: NA
Email: laxmiraju.k@maxisit.com
Website: www.maxisit.com