Article | Received 1 October 2025; Revised 24 November 2025; Accepted 16 December 2025; Published 31 December 2025
https://doi.org/10.55092/let20250009

From design to decommissioning: TAFES framework for responsible AI

Samia Loucif[1], Ravi Sharma[1,*], Nir Kshetri[2] and Arnob Zahid[3]

[1] College of Technological Innovation, Zayed University, Abu Dhabi 145534, United Arab Emirates
[2] Bryan School of Business and Economics, University of North Carolina, Greensboro 27412, USA
[3] Waikato Management School, University of Waikato, Hamilton 3240, New Zealand

* Correspondence author; E-mail: ravishankar.sharma@zu.ac.ae.

Highlights:

- AI systems are in critical need of governance and management in order to serve their intended requirements.
- An exploration of useable governance frameworks in both the public and private sector revealed five key constructs for determining responsible AI.
- Utilizing the TAFES guidelines across the AI life-cycle allows for the monitoring and control of AI applications.

Abstract: Artificial Intelligence (AI) systems demand comprehensive governance frameworks that ensure ethical, transparent, and accountable practices across their entire lifecycle. This paper presents the TAFES framework—comprising Transparency, Accountability, Fairness, Ethics, and Safety—as a theoretically grounded and practically implementable approach for responsible AI. Through systematic analysis of existing global frameworks and an applied healthcare use case (Nurse Linda), TAFES demonstrates how ethical principles can be operationalized through lifecycle processes spanning design, development, deployment, and decommissioning. Unlike regulatory or risk-centric models such as the NIST AI RMF or EU AI Act, TAFES integrates moral philosophy with engineering implementation to bridge the gap between ethical intent and practical execution.

Keywords: responsible AI implementation; AI governance framework; AI life-cycle management; AI regulation and policy; AI ethics; AI system decommissioning

1. Introduction

Ethics plays a vital role in the realm of technology, ensuring that advancements truly serve the interests of individuals and society. As technology continues to advance at a fast pace, it is essential to have ethical frameworks in place to guide our decision-making. Philosophers have offered different ways to think about this. For example, utilitarianism suggests that technology should be designed to increase

overall happiness and reduce harm. Deontological ethics, on the other hand, emphasizes sticking to moral rules like respecting privacy and ensuring fairness in how technology is used. Virtue ethics, which centers on the development of good character traits, encourages those involved in technology to act responsibly and ethically. Despite these different approaches, they all share a common goal, which is to guide actions that promote the well-being of individuals and societies, and to ensure that technology benefits people and society while minimizing any potential harm. This is what we mean by responsible Artificial Intelligence (RAI) in this paper.

The primary objective of this paper is to consider the question of whether we may develop a regulatory framework that would guide the responsible use of AI. Extraordinarily, the challenge of regulating AI-based systems has received the attention of scholars (Jobin *et al.* [1] and Sharma *et al.* [2]), industry (IBM [3], Microsoft [4,5] and Google [6,7]) and governments (Gong *et al.* [8], European Commission [9,10] and G20 [11]). When Luciano Floridi, the Chair of an EU initiative known as AI4People [12] prefaced a report with the revelation that "AI is not merely another utility that needs to be regulated only once it is mature. It is a powerful force that is reshaping our lives, our interactions and our environments", he probably thought he was stating the obvious. But to regulators and industry titans, it was not obvious until the present time that unless we govern AI, such systems may end up harming humanity.

There is no disagreement that AI has become an integral part of our daily lives, from virtual assistants to self-driving cars. As AI becomes more pervasive it is essential to ensure that its design and use align with moral values and ethical principles. In this paper, we examine the attributes of Fairness, Accountability, Transparency, Ethics, and Safety that are crucial for building AI that enables rather than inhibits users' techno-moral virtues. We propose a conceptual framework that incorporates these attributes along with privacy to engender "trust-worthy AI". We posit that such an approach to guide the adoption of AI for good will bring far greater benefits than the current competitive, free-for-all climate. Drawing on the vast experiences of Information Systems Audit and Control Association (ISACA) Control Objectives for Information Technologies (COBIT) [13], industry regulators could establish control objectives for techno-moral compliance in AI-based systems [13]. As a logical consequence, trustworthy audit procedures, credentialing of systems, and certification of practitioners with the benefit of industry insights will result in regulator-industry-stakeholder partnerships.

In context of generative AI (GenAI) implementation, governance is about implementation decisions that define requirements expectations, grant authority, assign responsibility and ensure compliance to stated specifications. Specifically, the responsible use of AI in industry and society must be governed by strategy (does AI align with our vision?), implementation maturity (are we capable of using AI effectively?), risk management (are risks manageable, *i.e.*, identifiable, mitigatable?), active monitoring (is AI overstepping its intended purpose?). Together, the governance of AI must assure safety, responsible use while preserving user privacy.

Figure 1 identifies the typical perspectives with which IT (and correspondingly, GenAI) systems and platforms are governed. It is clear that GenAI Governance would be even more complex, contextual and value-centric. There are governance issues to be monitored and controlled at each stage of the system life-cycle. Therefore, a manifesto for RAI must seek to address these contextual impacts or it would remain just that, a manifesto that reads well but with little utility.

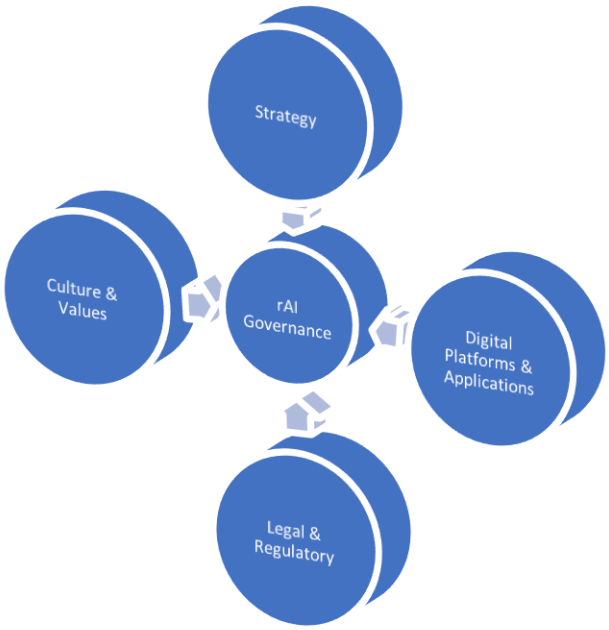

Figure 1. Scope of RAI governance.

In previous work, we have urged that techno-moral virtues be considered the foundational basis for applications of AI and data analytics (Sharma *et al.* [14]). In later work, we have discussed the regulatory imperatives for responsible AI (Sharma *et al.* [2]). In recent work, we have presented a framework for AI governance (Sharma *et al.* [15]). This article presents a more complete picture of the *TAFES* framework and how it might be applied across the AI life-cycle. It also argues that a regulatory framework for AI must address the question of what constitutes RAI. How might we design, develop, deploy, and when needed, decommission AI systems so that there is greater trust by users and safeguards against abuse by service providers? These questions are deep and significant, and in fact give rise to what may be termed paradoxical. The field research of Bilal *et al.* [16] and König *et al*. [17] have found, citizens or netizens are prepared to sacrifice privacy, transparency, and "involvement" (oversight?), in the interests of perceived convenience. Is this responsible though?

The rest of the article is structured as follows: Section 2 reviews diverse frameworks that have been proposed for the responsible use of AI. Section 3 presents the conceptual framework—our *TAFES* approach to designing and using AI for good. Section 4 discusses how the *TAFES* principles may be applied to a use-case of GenAI in healthcare. Section 5 discusses the limitations of the proposed regulatory framework and future work. Section 5 presents some limitations of the *TAFES* framework, highlighting its challenging integration across the AI lifecycle, current gaps, and our future directions. Finally, Section 6 concludes this article with RAI Manifesto (*TAFES*), outlining responsible AI implementations, stakeholder roles, current gaps, and future research needs.

2. Review of responsible AI frameworks

The formulation and implementation of RAI has been receiving increasing attention from key stakeholders, including academics, governments, technology firms, and professional societies. This section provides an overview of the numerous frameworks and guidelines proposed to address ethical considerations in AI and autonomous systems, ostensibly to protect society.

2.1. Academic perspectives

The academic community has been at the forefront of developing ethical frameworks for AI. Dignum [18,19], an expert in AI ethics and RAI, has contributed significantly to this field. She advocated for the integration of ethical, legal, and societal values throughout the AI lifecycle. Floridi *et al*. [20–22] emphasized the transformative impact of AI on society and the need for ethical considerations in AI governance. Recent developments in 2024 have seen increased focus on practical implementation, with the ACM FAccT conference emphasizing learning pathways for practitioners and the need for functional rather than aspirational frameworks [23]. Gebru's research [24] focused on ethical AI, bias in AI systems, and diversity in AI research.

Prunkl *et al*. [25] highlighted the importance of long-term effectiveness in AI ethics, advocating for greater transparency, clearer guidance, and strong incentives for meaningful engagement. They also stressed the significance of community-based governance and the wider responsibilities of the AI research community in this process.

Sinha *et al*. [26] addressed challenges in developing and deploying AI models and applications in industrial systems. They advocated for strategic measures to address data collection and management, algorithm selection and training, model interpretability, and fostering interdisciplinary collaboration. In the same vein, Sadek *et al*. [27] had pointed out the challenges of implementing RAI in their scoping review. Their recommendations called for the use of contextual understanding ahead of design implementation.

Raquib *et al*. [28] proposed a holistic Islamic virtue-based AI ethics framework grounded in the context of Islamic objectives (maqāṣid). They argued that this framework could be used to explore AI-related ethical problems more holistically due to its ontological base and rich tradition. In a more recent book, Khan [29] has crafted 11 Islamic Principles for the development and utilization of AI based on the concepts of stewardship, public interest, intention and morality. These are: (1) idea of Niyyah or intention, (2) dignity and respect for all creation, (3) justice and equity, (4) privacy, (5) transparency and accountability, (6) serving humanity and enriching lives, (7) prohibition of harm, (8) beneficial knowledge, (9) autonomy and free-will, (10) collaboration and inclusivity, and (11) intellectual property rights.

Cheng & Zhen [30] conducted a comparative analysis of AI ethics policies in China, the US, and the EU, focusing on privacy protection and AI ethics. Their work highlighted the global convergence of AI ethical standards, noting that China has been actively researching and introducing AI standards and ethics, largely aligning with Western democratic societies.

Corrêa NK *et al*. [31] conducted a comprehensive review of 200 AI governance guidelines worldwide, revealing significant convergence around core principles while highlighting implementation diversity across cultural contexts. This analysis provides strong empirical support for framework consolidation efforts.

Fairness is one of the core pillars in the *TAFES* framework. Kinney [32] discussed the evaluation of predictive algorithms in situations where concepts of accuracy and fairness conflict. The paper suggested a mathematical model that combines the two metrics with their corresponding weights; the weights reflect the value that stakeholders give to each metric. Using the COMPAS dataset, it showed that different choices in weighting fairness and accuracy can noticeably change judgments about the AI model performance.

2.2. Government policies and initiatives

Governments worldwide have recognized the need for comprehensive AI governance frameworks. The European Commission's Ethics Guidelines for Trustworthy AI (European Commission [33]) highlight fundamental rights, non-discrimination, and environmental and societal well-being as key principles. According to the EU, Trustworthy AI should be lawful, ethical, and robust, with the third imperative referring to adapting to changing social circumstances. The EU AI Act became operational in August 2024, establishing comprehensive regulatory requirements with significant penalties for non-compliance (European Commission [10]).

China's State Council's AI plan (FLIA [34]) declared the country's intention to actively participate in global governance of AI and set an ambitious timeline for doing so. Khanal *et al.* [35] suggested that China's global ambitions are grounded in local contextual factors. The plan aimed to "strengthen the study of major international common problems such as robot alienation and safety supervision, deepen international cooperation on AI laws and regulations, international rules and so on, and jointly cope with global challenges" (Cheng & Zhen [30]).

In India, the National Strategy on AI (NITI Aayog [36]) identified six system considerations for safe and RAI: (1) Understanding the AI system's functioning for safe and reliable deployment, (2) Post-deployment explanability, (3) Consistency across stakeholders, (4) Preventing incorrect decisions leading to exclusion from access to services or benefits, (5) Accountability of AI decisions, and (6) Privacy and Security risks. The 2024 India AI Mission allocated ₹10,371.92 crore for AI infrastructure and governance initiatives (IMPR [37]).

Emerging frameworks from Global South countries provide important perspectives often missing from Western-centric approaches. Brazil's AI Plan 2024–2028 emphasizes technological sovereignty and inclusive development (Brazilian Government [38]), while South Africa's AI Policy Framework positions the nation as an African leader in responsible AI governance (South African Department of Communications [39]).

The United Arab Emirates (UAE [40]) developed AI Ethics Guidelines through "a collaborative process in which all stakeholders were invited to be part of an ongoing dialogue." The UAE's guiding principles include fairness, accountability, transparency, traceability, explanability, robustness, human-centricity, sustainability, and privacy/data protection. Notably, the UAE framework explicitly states that "Humanity should retain the power to govern itself and make the final decision, with AI in an assisting role."

Singapore's approach to AI governance (PDPC [41]) covers 11 governance principles, including transparency, explanability, repeatability/reproducibility, safety, security, robustness, fairness, data governance, accountability, human agency and oversight, inclusive growth, and societal and environmental well-being. The 2024 updated framework specifically addresses generative AI with enhanced governance dimensions (Singapore IMDA [42]). Australia's Voluntary AI Safety Standard (DISR [43]) gives practical guidance to organisations on "how to safely and responsibly use and innovate with AI" to "ensure the development and deployment of AI systems in Australia in legitimate but high-risk settings is safe and can be relied on." Specifically, the standard consists of 10 voluntary guardrails that apply to all organisations throughout the AI supply chain, as shown in Box 1. They include transparency and accountability requirements across the supply chain. They also explain what developers and deployers

of AI systems can do so that Australians may benefit from AI while mitigating and managing the risks that AI may pose.

Box 1—Australia's Guardrails (reproduced from DISR [43], p. iv): (1) Establish, implement, and publish an accountability process including governance, internal capability and a strategy for regulatory compliance. (2) Establish and implement a risk management process to identify and mitigate risks. (3) Protect AI systems and implement data governance measures to manage data quality and provenance. (4) Test AI models and systems to evaluate model performance and monitor the system once deployed. (5) Enable human control or intervention in an AI system to achieve meaningful human oversight. (6) Inform end-users regarding AI-enabled decisions, interactions with AI and AI-generated content. (7) Establish processes for people impacted by AI systems to challenge use or outcomes. (8) Be transparent with other organisations across the AI supply chain about data, models and systems to help them effectively address risks. (9) Keep and maintain records to allow third parties to assess compliance with guardrails. (10) Engage your stakeholders and evaluate their needs and circumstances, with a focus on safety, diversity, inclusion and fairness.

The U.S. Government's NIST Strategy [44] emphasizes the importance of global AI standards, urgent standardization topics, diverse stakeholder participation, and international collaboration. Key challenges identified include risk management, addressing bias and fairness, ensuring security, and enhancing transparency in AI systems to build trust and reliability. The July 2024 release of the Generative AI Profile provides specific guidance for foundation model risks [44].

While other countries hold firm on imposing AI and ethics, the UK provides more freedom while holding back prescriptive early laws in exchange for voluntary codes of practice, sectoral guidance, and light-touch regulations. Its White Paper on AI and National AI Strategy concentrate on building support for innovation while integrating key ethics like transparency, fairness, accountability. Fostering existing regulators and industry–government partnership building is preferred, in exchange for central drivers, harsh controls. Entities like DSIT and AI Safety Institute prefer testing, monitoring, and ethics matching, not killing innovation (Department for Science, Innovation and Technology [45], UK Government [46]).

2.3. Global technology firms' initiatives

Major technology firms have also developed their own approaches for RAI. Microsoft's Approach (Microsoft [18,30,45]) to AI investment focuses on transparency, accountability, fairness, inclusiveness, reliability, safety, privacy, and security as foundational values guiding their AI product development. Their 2024 RAI Transparency Report documents deployment of 30+ responsible AI tools with comprehensive governance structures (Microsoft [18]). IBM [3], an early adopter of the notion of ethics for AI, has an entire AI Ethics constituting ethics boards and identifying risk associated with training GenAI models. The company AI Ethics Project Office supports all these initiatives, serving as a liaison between governance roles, supporting implementation of technology ethics priorities, helping establish AI Ethics Board agendas and ensuring the board is kept up to date on industry trends and company.

In India, Lohchab [47] reports that major IT firms have constituted internal review boards (similar to Institutional Review Boards) to ensure ethical deployment of AI, particularly in response to concerns surrounding GenAI. These measures are intended to address issues of data privacy, sovereignty, bias, and hallucinations.

2.4. Professional societies and collaborative initiatives

Professional societies and collaborative initiatives have played a crucial role in developing comprehensive frameworks for RAI. The IEEE Global Initiative for Ethical Considerations in AI and Autonomous Systems (IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems [48]) emphasizes the importance of transparency, accountability, and privacy in AI design. The Asilomar AI Principles (Future of Life Institute [49]), created by a group of German AI researchers and stakeholders, focus on values such as transparency, fairness, and human control. The Information Systems Audit and Control Association (ISACA [50]) has formulated 'Considerations for Implementing a GenAI Policy,' which outlines AI Acceptable Usage Policy (AUP) guidelines. These guidelines address key areas such as secure AI systems, ethical AI principles, data handling and training guidelines, transparency and attribution, legal and compliance requirements, and human oversight.

In a pan-European initiative, AI4People [12] was launched by the European Commission and chaired by Professor Luciano Floridi to develop principles, policies, and practices for building a "good AI society." The initiative resulted in the AI4People's 7 AI Global Frameworks, focusing on seven strategic sectors for the deployment of ethical AI.

2.5. Synthesis of perspectives

Notwithstanding the proliferation of over 80 publicly accessible RAI frameworks which promote societal values, Sadek and co-workers [27] lament that "the techno-centric toolkits that 'hungry methodologists' created to solve RAI, fail to account for the pluralistic requirements of a specific application context." To address this, they proposed a toolkit incorporating worksheets and cards for discussions and eliciting explicit statements of participants' preferences, used data types, essential values, and its application context. But this has not received significant adoption in industry nor society. Why? We may deduce from the above review that, while numerous frameworks and guidelines have been proposed by various stakeholders to promote RAI, challenges remain in implementing these principles effectively across diverse contexts and ensuring they adequately protect society. As Bughin [51] insightfully analyzed, there is a disconnect between doing and saying, precisely because there were insufficient means of checking and verifying compliance.

Therefore, three fundamental question come to fore:

(1) What constitutes acceptable AI adoption over decisions impacting human lives?

(2) Is there a universal view of ethics and living well with AI?

(3) Why is there no agreement on who is accountable for an AI outcome, users themselves, the service providers or the regulators?

The ongoing dialogue and collaboration between academics, governments, technology firms, and professional societies will be crucial in refining and implementing these frameworks to guide the ethical development and deployment of AI technologies.

Table 1 categorizes the key issues, challenges, and resolutions in designing RAI for the major stakeholder groups mentioned. The table is structured to provide a clear overview of how different stakeholders approach RAI. Five key observations may be made from the analysis of Table 1:

(1) There is a significant overlap in issues across stakeholder groups, particularly regarding transparency, fairness, and privacy.

(2) Governments and regulators are focusing on creating comprehensive frameworks and guidelines, often with a global perspective.

(3) Academics and scholars are delving deeper into specific aspects of AI ethics and proposing novel concepts and frameworks.

(4) Technology firms are working on practical implementations of RAI principles, often creating their own guidelines or review processes.

(5) The needs and concerns of diverse users are being considered by all other stakeholder groups, emphasizing the importance of human-centric AI design that will promote its successful deployment.

Table 1. RAI governance initiatives across stakeholders.

Stakeholder Group	Issues	Challenges	Resolutions
Governments & Regulators	• Ethical deployment of AI • Data privacy and sovereignty • Fairness and non-discrimination • Transparency and accountability • Globally adopted standards for governance of AI	• Balancing innovation with regulation • Addressing cross-border AI governance • Keeping pace with rapid AI advancements	• China's State Council AI plan for global governance participation • NITI Aayog's 6 system considerations for safe AI • EU's Ethics Guidelines for Trustworthy AI • UAE's AI Ethics Guidelines with 9 principles • Singapore's 11 governance principles • U.S. NIST Strategy for AI standards • UK pro-innovation approach to AI regulation
Academics & Scholars	• AI ethics and RAI • Bias in AI systems • Diversity in AI research • Long-term implications of AI • Transparency and explainability	• Ensuring long-term effectiveness of AI ethics • Implementing meaningful transparency • Addressing pluralistic requirements in specific contexts	• Dignum's work on AI ethics and RAI • Floridi's emphasis on ethical considerations in AI governance • Gebru's research on bias and diversity • Rajan *et al.*'s Islamic virtue-based AI ethics framework • Prunki *et al.*'s call for community-based governance • Sloane *et al.*'s concept of contextual transparency • Sinha *et al.*'s strategic measures for AI in industrial systems
International Technology Firms	• RAI development and deployment • Addressing user concerns • Ensuring ethical use of AI products	• Balancing innovation with ethical considerations • Managing risks associated with AI deployment • Ensuring consistency across global operations	• Microsoft, IBM, as examples, their approach focusing on transparency, accountability, fairness, *etc*. • Indian IT firms establishing internal review boards • ISACA's AI Acceptable Usage Policy (AUP) guidelines
Diverse Users & Societies	• Privacy protection • Fairness and non-discrimination • Transparency in AI decision-making • Human agency and control	• Understanding AI-driven decisions • Protecting against misuse or harm from AI systems • Ensuring AI benefits are inclusive	• Calls for human-centric AI design • Demands for explainable AI • Emphasis on inclusive growth and societal well-being in AI development • Auditing, ISACA

3. *TAFES* framework

3.1. Framework derivation methodology

The *TAFES* framework emerges from a systematic comparative analysis of existing RAI frameworks conducted between 2023 and 2024. Our approach involved taking the following steps:

Step 1: Framework Collection and Analysis. We identified and analyzed 47 comprehensive AI governance frameworks from 24 countries and 12 major technology companies. Frameworks were selected based on comprehensiveness, implementation specificity, and stakeholder representation. Frameworks that lacked explicit lifecycle coverage, operational guidance, or multi-stakeholder validation were excluded to ensure methodological rigor and reproducibility.

Step 2: Principle Extraction and Frequency Analysis. Through systematic content analysis using a premium subscription of Claude Sonnet 4.5, we identified 17 core principles appearing across frameworks. Frequency analysis revealed the most common principles: (1) Transparency: 92% of frameworks; (2) Safety/Security: 94% of frameworks; (3) Accountability: 88% of frameworks; (4) Fairness: 85% of frameworks; (5) Ethics: 79% of frameworks.

Step 3: Synthesis and Validation. The five *TAFES* principles were selected based on: (a) frequency of appearance across frameworks, (b) operational feasibility for implementation, (c) comprehensive coverage of responsible AI concerns, and (d) expert validation through consultation as a research team and feedback from AI subject matter experts at seminars or symposiums.

Table 2 presents a comparative analysis showing how *TAFES* compares to existing frameworks across dimensions of Principle Coverage, Lifecycle Integration, Implementation Specificity, and Global Applicability.

Table 2. Comparison of *TAFES* principles against select frameworks.

Framework	Principle Coverage	Lifecycle Integration	Implementation Specificity	Global Applicability
EU AI Act	Regulatory focus on high-risk systems	Primarily deployment	High (legal requirements)	EU + extraterritorial
NIST AI RMF	Comprehensive risk management	Full lifecycle	Medium (voluntary guidance)	Global adoption
Microsoft RAI	Technical implementation focus	Development + deployment	High (technical tools)	Corporate context
TAFES	Five core principles	Complete (design to decommission)	High (practical guidance)	Global + cross-cultural

3.2. Theoretical foundations of the TAFES framework

We claim that the *TAFES* framework is grounded in a synthesis of ethical traditions and governance theories, offering a coherent normative architecture for responsible AI. Each principle (*i.e.*, Transparency, Accountability, Fairness, Ethics, and Safety) draws from distinct philosophical and organizational foundations, enabling both conceptual clarity and practical implementation. More specifically:

Transparency is rooted in Enlightenment rationalism and virtue ethics, emphasizing intelligibility, scrutiny, and autonomy. It reflects the moral imperative to ensure that AI decisions are understandable

to those affected. In governance theory, transparency aligns with public accountability and procedural justice, operationalized through explanability protocols and verifiability mechanisms.

Accountability emerges from deontological ethics and moral responsibility, requiring that actors answer for their decisions across the AI lifecycle. Organizationally, it resonates with stewardship and agency theory, clarifying roles and obligations. *TAFES* embeds accountability through traceable decision records and structured audit mechanisms, including RACI-style matrices that delineate Responsible, Accountable, Consulted, and Informed roles.

Fairness is informed by Rawlsian justice and the social contract tradition, mandating equitable treatment and non-discrimination. In AI contexts, this principle translates into bias mitigation, demographic representativeness, and outcome equity. *TAFES* advances fairness beyond normative aspiration by prescribing measurable metrics such as demographic parity and equalized odds, integrated throughout design, development, and deployment stages.

Ethics serves as the integrative dimension, drawing on virtue ethics and care ethics to promote deliberative reflection and moral pluralism. Rather than treating ethics as a post hoc compliance exercise, *TAFES* embeds ethical reasoning within lifecycle processes, ensuring that cultural and institutional contexts inform design and governance decisions.

Safety synthesizes utilitarian principles and systems-engineering perspectives, prioritizing harm minimization and risk mitigation. It incorporates the bioethical tenet of nonmaleficence ("do no harm") and extends safety beyond technical constraints to include procedural and moral safeguards across the AI lifecycle.

Unlike frameworks such as the NIST AI RMF, which emphasize risk management, or the EU AI Act, which centers on regulatory compliance, *TAFES* unifies ethical theory with implementable governance. Its philosophical coherence and operational specificity position it as a bridge between aspirational ideals and actionable standards.

Contextual Adaptability is a defining strength of *TAFES*. The framework allows for calibration across domains and jurisdictions: Safety and Ethics are prioritized in healthcare, while Transparency and Accountability are critical in finance and public administration. In low-resource settings, Fairness and implement-ability gain prominence. This flexibility ensures that *TAFES* remains globally applicable without compromising its moral integrity.

TAFES is designed for contextual calibration. The emphasis on specific principles can vary by domain or societal setting: for instance, Safety and Ethics predominate in healthcare, while Transparency and Accountability are more critical in financial and public sector applications. In low-resource contexts, Fairness and implementability assume greater importance. This flexibility allows regulators and organizations to tailor *TAFES* implementation without compromising its core moral integrity. The *TAFES* principles also serve as a foundation for developing and deploying AI systems in a responsible manner, ensuring they align with ethical standards and contribute positively to society while minimizing potential harm. More specifically, the *TAFES* framework provides an opportunity to act on the call for by the National Institute of Standards and Technology of the United for global engagement in formulating standards with multi-stakeholders for the alignment, development, and adoption of AI. According to Dunietz *et al*. [52], this would require processes that are "open, transparent, and consensus-driven" across the lifecycle. It is therefore clear that transparency must be front and center of any approach to RAI systems.

Figure 2 captures the synergy among the enabling affordances of RAI. In defining these affordances (system characteristics or features which may be perceived or actual, enabling or inhibiting), we may use the lens of the AI Systems Lifecycle to further delineate the following distinct phases.

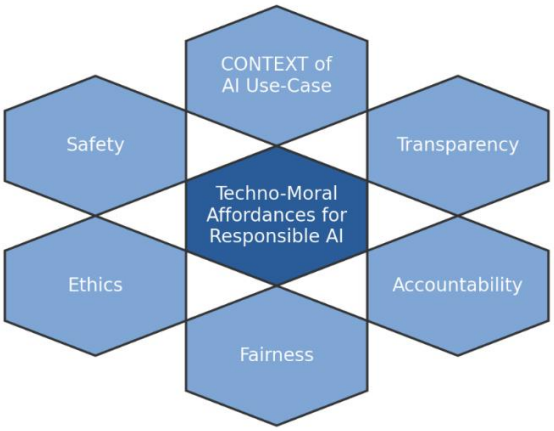

Figure 2. Techno-Moral Framework of RAI.

Transparency requires that the decision-making process of AI systems should be open and understandable to humans. This helps to build trust and ensures that AI is not used to harm individuals or groups.

Accountability is another essential component of responsibility. Those responsible for developing and deploying AI systems should be held accountable for their actions, ensuring that the AI is used in an ethical manner and does not cause harm.

Fairness in AI means that algorithms and systems should not discriminate against individuals or groups based on their race, gender, or other characteristics. Ensuring fairness in AI systems is a challenge due to the inherent biases and lack of diverse representation in data sets.

Ethics in AI systems should consider broader societal and environmental impacts beyond economic benefits, aligning with societal values and promoting the public good.

Safety (incorporating Security, Privacy, Data Protection) is integral to AI systems and should respect individuals' privacy rights, comply with data protection regulations, and ensure data is collected, used, and stored lawfully, with safeguards against unauthorized access or use to prevent harm to users.

We do not claim novelty in proposing such a framework distilled from a background review. However, when benchmarked with a more static framework such as the one used by Sinha *et al.* [26], the *TAFES* model, presents a set of affordances that may be distinctly specified across the lifecycle phases of AI systems. Moreover, the utility of the *TAFES* framework lies in its articulation of AI implementation rather than provide a checklist of features. For example, Taylor [53] has questioned whether existing explainable AI techniques can indeed close the responsibility gap and identifies a number of significant limits to their ability to do so.

Consistent with the traditional systems development lifecycle, implementing an AI system begins with requirement analysis and planning to elicit the requirements such as the decisions to be made, data to be used, stakeholders involved, goals and performance metrics to be set, and an examination of benefits and costs. Following such a classical feasibility analysis, we propose the following phases for the purpose of developing our framework:

(a) Design: The phase where the conceptualization and planning of the AI system's structure, components, and functionalities take place to meet specified requirements and objectives, focusing on specification of functional and non-functional requirements, architecture; key stakeholders: technology providers, lead users.

(b) Develop: The stage where the actual implementation, coding, and testing of the AI system occur based on the design specifications to create a functional prototype or product, translating design specifications into functional software; key stakeholders: technology providers, regulators. Then testing the system to validate safe and responsible use.

(c) Deploy: The process of integrating the developed AI system into its operational environment, ensuring compatibility, performance, and functionality within the intended context, including testing, integration, and user training, and key stakeholders: technology providers, users.

(d) Decommission: The systematic shutdown, removal, or replacement of the AI system at the end of its lifecycle when its net utility is zero or negative, including data disposal and transitioning to newer technologies or processes, involving the retirement and removal of AI systems from service, ensuring data security and ethical considerations are addressed during the shutdown process; key stakeholders: regulators; users.

Juxtaposing *TAFES* attributes with the AI Lifecycle, the part below (*TAFES* affordances across AI lifecycle) outlines some of the key techno-moral affordances of RAI as they apply across design, develop, deploy, and decommission phases.

3.3. TAFES affordances across AI lifecycle

3.3.1. TRANSPARENCY

AI systems should operate transparently, enabling users and stakeholders to understand their functioning and decision-making processes to identify potential biases or issues.

(a) Design: Plan for transparency in design by documenting decision-making processes and system operations with UML diagrams, algorithms, training datasets and reporting measures of bias, accountability, and transparency. Design the system with built-in explanability features, such as methods to highlight which features are most influential in the model's predictions and visual tools to help users understand model decisions and feature contributions. User = friendly interfaces are essential to present the explanations.

(b) Development: During development, implement both the AI system and the chosen explanability methods, incorporating transparency features that allow stakeholders to understand the limitations and functions of the AI system.

(c) Deployment: Provide stakeholders with access as required to information about deployed AI system operations and decision-making processes. Monitor how effectively these transparency features help stakeholders understand the system decisions and limitations. Collect feedback to improve the explanations as needed.

(d) Decommissioning: Maintain transparency during decommissioning by documenting the process and informing stakeholders about reasons for shutdown and ensuing procedures.

3.3.2. ACCOUNTABILITY

AI systems should be accountable for their actions, and those involved in their development and deployment should be held responsible for any negative impacts.

(a) Design: Establish clear accountability structures during design to ensure the who-what-when of responsible decision-making. Develop valid explanations when required.
(b) Development: Incorporate accountability measures into development to track decisions made by the AI system and provide a mechanism for feedback from lead users on the merits of explanations.
(c) Deployment: Ensure operational accountability mechanisms such as explanations during deployment to address issues and hold stakeholders responsible with audit trails, tracking logs and exceptions reports.
(d) Decommissioning: Define accountability protocols for decommissioning when accountability standards are not met to manage responsible shutdown and disposal of data logs.

3.3.3. FAIRNESS

AI systems should be designed and implemented in a way that are fair to all individuals and groups, avoiding perpetuating or amplifying existing biases or discrimination.

(a) Design: Ensure AI systems are designed to be fair and unbiased, considering diverse perspectives and avoiding discrimination in the choice of training datasets, algorithms and measures of accuracy.
(b) Development: Implement measures to monitor and address biases during development (specifically data preprocessing and feature selection, training the model then validation testing) to promote fairness.
(c) Deployment: Verify that deployed AI systems maintain fairness in real-world applications and address emerging biases promptly by auditing accuracy and providing remedies.
(d) Decommissioning: Safely decommission AI systems when they fail the fairness thresholds, considering fairness impacts and handling data ethically.

3.3.4. ETHICS

AI systems should consider broader societal and environmental impacts beyond economic benefits, aligning with societal values and promoting the public good.

(a) Design: Integrate ethical considerations into design to align AI systems with societal values, and ethical standards in accordance with legal-regulatory directives.
(b) Development: Uphold ethical principles throughout development lifecycle, incorporating societal norms and socio-economic justice.
(c) Deployment: Ensure deployed AI systems adhere to ethical guidelines (both perceived and as well as realized) and promote the public good.
(d) Decommissioning: Monitor and consider ethical violations prior to decommissioning to retire AI systems responsibly and address ethical concerns.

3.3.5. SAFETY (including security, privacy, data protection)

AI systems should respect individuals' privacy rights, comply with data protection regulations, and ensure data is collected, used, and stored lawfully, with safeguards against unauthorized access or use to prevent harm to users.

(a) Design: Incorporate safety, security, privacy, and data protection functionalities in design in strict compliance with legal and industry requirements to build trust and compliance.
(b) Development: Implement robust safety and security measures and mechanisms during development to protect data and ensure system integrity in terms of compliance with laws and regulations.
(c) Deployment: Maintain focus on the monitoring and control of safety, security, privacy, and data protection during deployment to safeguard users and information.
(d) Decommissioning: Address safety, security, privacy, and data protection concerns during operations and initiate decommissioning to retire AI systems when safeguards are not met; ensuring that sensitive data are handled in compliance with legal and ethical obligations.

3.4. Annotated discussion

The commentary in section 3.3 reveals the need for a multi-stakeholder perspective to be applied socio-technically across the stages of design, development, deployment, and decommissioning, albeit in summary format. For example, how might we build enabling and realizable affordances (resulting in beneficial outcomes) of fairness throughout the lifecycle? At the design phase, recognize that fairness requires a minimization of bias in the model, training dataset and outcome. In the development phase, specific efforts to counter bias must be taken, transparently. During deployment, outcomes must be monitored against effective measures that detect bias. If bias exceeds a stated threshold, service providers must be accountable to stakeholders who would decide to decommission the system as they see fit.

To drill deeper, how might AI service providers factor in accountability functionalities during the design phase? Some guiding principles drive our manifesto for RAI: (1) Establish clear accountability structures in AI system design, such as experts to be consulted, regulations to be met, *etc.* (2) Incorporate accountability measures into the development phase, such as logs and trails for decisions made and their outcomes. (3) Establish clear lines of responsibility within the development team so that critical features of responsibility do not fall between the cracks. (4) Plan regular assessments and evaluations of the AI system's performance against ethical standards. (5) Explicitly acknowledge compliance with regulatory frameworks, guidelines, and obligations. (6) Provide avenues for recourse and redress in case of malfunctions or ethical violations.

In the next section, we use a hypothetical but realistic case study of GenAI in Personalized Health Care as a Service (PHCaaS) to illustrate the utility of the *TAFES* principles in the implementation of RAI.

4. Use-case of TAFES in personalized health-care

In this section, we apply the *TAFES* principles to a hypothetical but realistic use case of AI. Our approach draws from how the Australian Voluntary AI Safety Standard was applied to realistic examples to illustrate how the 10 guardrails might be applied across the lifecycle. As the DISR [12] noted, “These examples are not intended to represent a comprehensive application of all relevant guardrails, responsibilities or other legal obligations that may be relevant for the specified use cases, but to explore how organizations may use particular guardrails as part of their overall approach to deploying AI systems.”

Taking a similar approach, we may now consider the design, development, deployment, and decommissioning of an application of GenAI in Personalized Healthcare as depicted by the recorded skit of Nurse Linda, a Generative AI agent developed by https://hippocraticai.com/. This agent is “designed to

follow up with a patient admitted and discharged for Congestive Heart Failure (CHF).” In this use case scenario, we examine the application of the *TAFES* (Transparency, Accountability, Fairness, Ethics, and Safety) principles to “Nurse Linda,” the AI Clinical Assistant. By analyzing the implementation of *TAFES* principles across the AI system lifecycle, we aim to demonstrate how RAI can be achieved in healthcare applications, concerning the growing impulse for global AI regulation (Jobin *et al.* [1]).

4.1. Nurse Linda the AI chatbot—technical implementation

Nurse Linda is an LLM-trained AI chatbot developed to provide follow-up care for patients discharged after CHF treatment. Its primary functions include monitoring patient symptoms, checking medication adherence, providing guidance on lifestyle adjustments, and following up with patients for 30 days or more post-discharge, adhering to the CHF Follow-Up Call Checklist protocol. Figure 3 below illustrates the design functionality of Nurse Linda in the post-discharge care process.

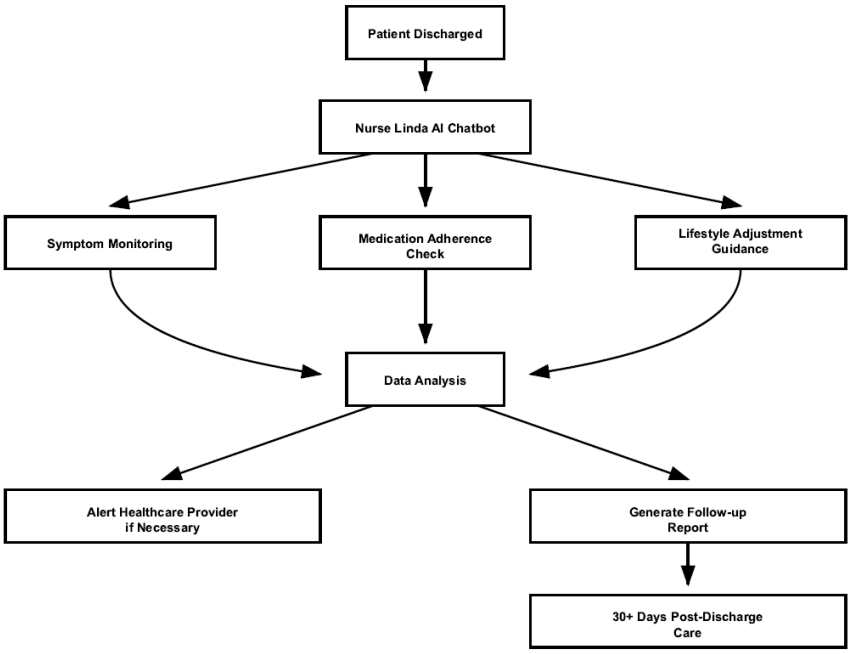

Figure 3. Post-discharge care process flow of Nurse Linda.

4.1.1. Patient discharge process flow

(1) Patient Discharged→triggers the system
(2) Nurse Linda AI Chatbot→central hub that initiates three parallel monitoring processes:
- Symptom Monitoring
- Medication Adherence Check
- Lifestyle Adjustment Guidance

(3) All three processes feed into Data Analysis
(4) Data Analysis leads to two outcomes:
- Alert Healthcare Provider if Necessary
- Generate Follow-up Report

(5) The Follow-up Report leads to 30+ Days Post-Discharge Care

4.1.2. Patient authentication and access

Patient Contact→Multi-factor Authentication→Medical Record Verification→AI Consent Process→Bias Detection Check→Clinical Protocol Alignment→Interaction Initiation→Continuous Safety Monitoring

4.1.3. Bias detection implementation: the system implements real-time bias monitoring across multiple dimensions

- Demographic bias detection (age, gender, race, insurance status)
- Clinical bias assessment (symptom interpretation, treatment recommendations)
- Response quality monitoring across patient populations
- Automated fairness metric calculation (demographic parity, equalized odds)

4.1.4. Explanability mechanism which comprises a three-layer transparency approach

(1) Patient Layer: Simple explanations ("I'm recommending you contact your doctor because your symptoms suggest...")

(2) Clinician Layer: Detailed reasoning with confidence scores and evidence citations

(3) Audit Layer: Complete algorithmic transparency with SHAP values and decision pathways

4.1.5. Technical architecture

Figure 4 is a detailed use case diagram which illustrates the complex interactions of Nurse Linda with key stakeholders. Taken together with Figure 3, it is clear that RAI safeguards are essential requirements of developing such an AI Agent in the healthcare ecosystem. The diagrams also underscore the critical role of RAI principles in managing interactions and process checks. It includes AI technologies such as LLMs, Voice Chatbot, and different data models as intermediaries to facilitate and improve patient-physician communications.

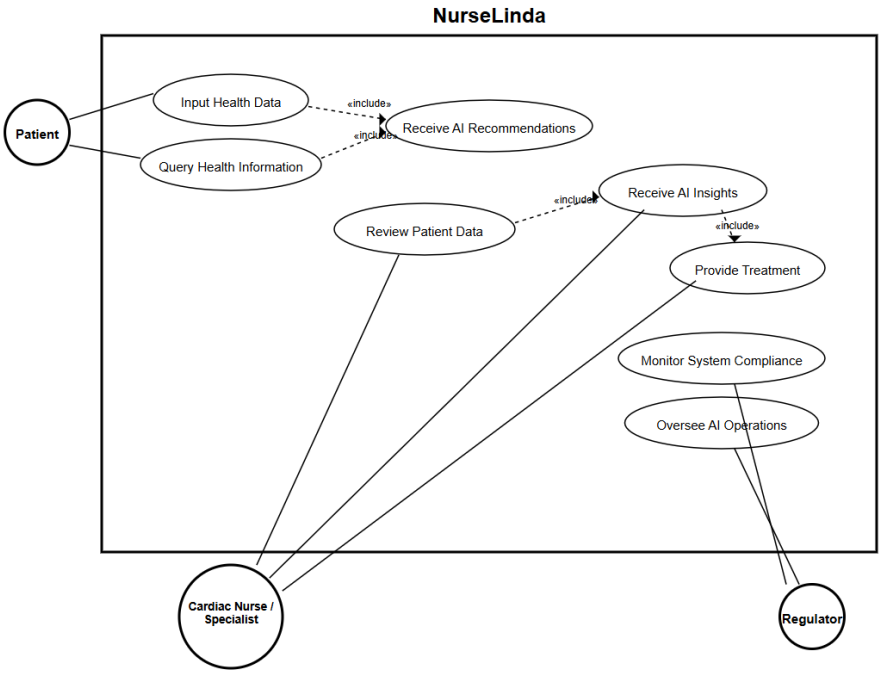

Figure 4. Use-case diagram of stakeholder interactions with Nurse Linda.

More specifically, Figure 4 illustrates the enhanced functionalities of Nurse Linda, the AI Chatbot and how various actors, including the Cardiac Nurse/Specialist, interact with it through comprehensive *TAFES* implementation. The utility of the *TAFES* framework is given by the extent to which Figures 3 and 4 align with the *TAFES* principles, showcasing the transparent and accountable nature of the AI chatbot's interactions with patients, healthcare professionals, and assuaging regulators' and service providers' ethics and safety considerations in the context of cardiac care.

4.1.6. Design rules

Following on from the requirements specifications and technical architecture in sections 4.1.4 and 4.1.5, the following design rules may be derived.

(1) Actors and Interactions:

- Patient (with enhanced authentication and consent processes)
- Cardiac Nurse/Specialist (with explanability dashboards)
- Regulator (with comprehensive audit access)
- Bias Monitor (new role for fairness oversight)
- Safety Officer (new role for continuous monitoring)

(2) Use Case Interactions:

- Input Health Data (with bias detection)
- Query Health Information (with explanability)
- Receive AI Recommendations (with transparency layers)
- Review Patient Data (with audit trails)
- Receive AI Insights (with confidence scoring)
- Provide Treatment (with human oversight)
- Monitor System Compliance (with automated alerts)
- Oversee AI Operations (with safety dashboards)
- Audit System Performance (with comprehensive logging)
- Manage Bias Detection (with real-time monitoring)

(3) Relationships Among Entities:

- Solid lines represent direct interactions between actors and use cases with TAFES implementation
- Dotted lines represent dependencies or information flow between use cases with audit trails
- New monitoring flows for bias detection, safety oversight, and regulatory compliance

(4) Process Logic for Technical Implementation:

(a) Patient Authentication Process:

```
def authenticate_patient(patient_id, biometric_data, access_code):
    # Multi-factor authentication
    identity_verified = verify_identity(patient_id, biometric_data)
    access_granted = validate_access_code(access_code)
    consent_status = check_ai_consent(patient_id)
```

```
    if identity_verified and access_granted and consent_status:
        return initiate_ai_session(patient_id)
    else:
        return redirect_to_human_support()
```

(b) Bias Detection Module:

```
def monitor_bias_metrics(patient_data, system_response):
    demographics = extract_demographics(patient_data)

    # Calculate fairness metrics
    demographic_parity = calculate_demographic_parity(responses_by_group)
    equalized_odds = calculate_equalized_odds(responses_by_group)

    # Check thresholds
    if demographic_parity < PARITY_THRESHOLD:
        trigger_bias_alert("demographic_parity_violation")
        apply_bias_mitigation(system_response, demographics)

    log_bias_metrics(demographic_parity, equalized_odds)
    return adjusted_response
```

(c) Transparency and Explanability Monitoring:

```
def generate_explanation(prediction, patient_profile, audience_type):
    # Calculate SHAP values for feature importance
    shap_values = calculate_shap_values(prediction, model_inputs)

    if audience_type == "patient":
        explanation = create_patient_friendly_explanation(
            prediction, key_factors=shap_values[:3]
        )
    elif audience_type == "clinician":
        explanation = create_clinical_explanation(
            prediction, shap_values, confidence_score, alternatives
        )
    elif audience_type == "auditor":
        explanation = create_technical_explanation(
            prediction, shap_values, model_metadata, training_data_info
        )
    log_explanation_interaction(patient_profile, explanation, audience_type)
    return explanation
```

The enhanced use case scenario is now detailed below, highlighting the main actors and system components with *TAFES* integration. This scenario will be used in our subsequent fit-for-purpose analysis of the *TAFES* parameters for RAI development.

(5) Additional features and characteristics of system components with *TAFES* dimensions:

- Patient: The prime user with enhanced authentication, consent management, and explanation access
- AI Interface: Enhanced with bias detection, explanability features, and safety monitoring
- LLMs/Chatbot: Equipped with fairness algorithms, transparency mechanisms, and safety constraints
- Bias Detection Module: Real-time monitoring of demographic and clinical bias with automated alerts
- Explanability Engine: Multi-layer explanation generation for patients, clinicians, and auditors
- Data Models: Enhanced with bias mitigation, audit trails, and safety validation
- AI Decision Support: Integrated bias checking, explanation generation, and safety assessment
- Doctor: Enhanced dashboard with AI explanations, bias alerts, and safety notifications
- Wearable Devices: Data validation and bias detection for sensor inputs
- Medical Records: Comprehensive audit trails and bias impact assessment
- Medical Knowledge Base: Validated guidelines with bias assessment and safety protocols
- Regulatory Compliance: Real-time monitoring with automated reporting and audit support
- Human Oversight: Enhanced safety monitoring with automated escalation procedures
- Audit Trail System: Comprehensive logging of all interactions, decisions, and bias metrics
- Safety Monitoring Dashboard: Real-time performance tracking with automated alerts

4.2. Use-case application of TAFES principles

To ensure comprehensive, responsible design, development, deployment, and potential decommissioning of Nurse Linda, we provide detailed implementation of the *TAFES* principles across each phase of the AI system lifecycle, building upon the framework proposed by Sharma *et al.* [14] with enhanced technical specifications. In the discussion that follows, Table 3 below provide a comprehensive summary of the application of *TAFES* Principles across all lifecycle phases using the specific implementation requirements of Nurse Linda, the AI Agent. It is a matrix that illustrates specific technical requirements, measurement criteria, and validation procedures for each *TAFES* principle across the lifecycle phases. Although presented as distinct tables for each life-cycle phase, it is clearly intended as a cyclic and iterative process. For example, if the design requirements cannot be translated into development specifications, the process flows back to redesigning the *TAFES* condition.

Table 3. Life-cycle considerations for RAI.

(a) Design Phase AI Implementation				
Transparency Design Requirements	Plan for multi-layer explanation architecture addresses to patients, clinicians, and auditors	Design user interfaces clearly indicating AI involvement with explanation access	Establish explanation quality metrics and validation procedures	Document decision-making processes with UML diagrams and algorithm specifications
Accountability Design Requirements	Define comprehensive responsibility matrix for all system components	Establish audit trail architecture with immutable logging capabilities	Plan escalation procedures with automated trigger mechanisms	Design incident response workflows with clear stakeholder notification protocols
Fairness Design Requirements	Plan for diverse training data with demographic representation analysis	Design bias detection algorithms with multiple fairness metrics	Establish fairness threshold definitions and violation response procedures	Plan for bias mitigation strategies across pre-processing, in-processing, and post-processing
Ethics Design Requirements	Integrate medical ethics guidelines into system design specifications	Plan patient consent mechanisms with clear AI capability disclosure	Design cultural sensitivity features and communication adaptation	Establish ethics review board oversight and approval processes
Safety Design Requirements	Plan comprehensive safety monitoring architecture with real-time alerts	Design fail-safe mechanisms and emergency decommissioning procedures	Establish data protection and privacy safeguards with encryption protocols	Plan regulatory compliance tracking and reporting mechanisms

Table 3. *Cont.*

(b) Development Phase Implementation				
Transparency Development	Implement SHAP-based explanability with audience-specific output formatting	Develop explanation interfaces with interactive exploration capabilities	Create explanation quality assessment tools with user feedback collection	Implement logging mechanisms for all explanation interactions
Accountability Development	Implement comprehensive audit trail system with searchable capabilities	Develop responsibility tracking with automated role assignment	Create incident detection and response automation with escalation protocols	Implement performance monitoring with accountability metric calculation
Fairness Development	Implement real-time bias detection with automated fairness metric calculation	Develop bias mitigation algorithms with effectiveness validation	Create fairness monitoring dashboards with demographic breakdown	Implement bias correction procedures with outcome tracking
Ethics Development	Implement informed consent management with dynamic updates	Develop ethics compliance checking with guideline validation	Create cultural adaptation features with patient preference integration	Implement ethics violation detection with automated reporting
Safety Development	Implement safety monitoring with automated threshold detection	Develop emergency procedures with immediate human escalation	Create security safeguards with intrusion detection and response	Implement data protection with encryption and access control
(c) Deployment Phase Implementation				
Transparency Deployment	Provide clear AI disclosure to all patients with explanation access	Offer multi-level explanations with user preference adaptation	Monitor explanation effectiveness with user feedback collection	Maintain transparency documentation with regular updates
Accountability Deployment	Establish clear escalation procedures with 24/7 monitoring	Implement ongoing monitoring and auditing with automated reporting	Provide accountability dashboard access for stakeholders	Maintain incident response readiness with regular testing
Fairness Deployment	Monitor bias metrics continuously with automated alert generation	Address emerging biases with immediate mitigation procedures	Verify equitable access and outcomes across patient populations	Provide fairness reporting with demographic impact analysis
Ethics Deployment	Continuously evaluate adherence to medical ethics with regular assessment	Maintain patient autonomy with clear opt-out mechanisms	Monitor cultural sensitivity with patient feedback integration	Provide ethics compliance reporting with violation tracking
Safety Deployment	Implement real-time safety monitoring with automated intervention	Ensure robust data protection with continuous security assessment	Maintain emergency response capabilities with regular testing	Provide safety performance reporting with improvement recommendations
(d) Decommissioning Phase Implementation				
Transparency Decommissioning	Clearly communicate decommissioning reasons to all stakeholders	Provide comprehensive final performance and explanation quality reports	Document lessons learned with recommendation for improvement of transparency	Ensure knowledge transfer with detailed documentation handover
Accountability Decommissioning	Provide timeline and process transparency with stakeholder communication	Complete final audit with comprehensive performance assessment	Ensure proper handover with responsibility transfer documentation	Archive accountability records with long-term access procedures
Fairness Decommissioning	Ensure equitable transition plans for all affected patient populations	Complete final bias assessment with demographic impact analysis	Provide fairness performance summary with recommendation for improvement	Ensure non-discriminatory decommissioning process across all groups
Ethics Decommissioning	Conduct thorough ethical review of AI performance and impact	Ensure patient notification with clear communication about alternatives	Complete ethics compliance final assessment with violation summary	Provide ethical lessons learned with future implementation guidance
Safety Decommissioning	Ensure equitable transition plans with minimal service disruption for patients	Complete comprehensive safety assessment with incident analysis	Implement secure data preservation with privacy protection measures	Provide safety performance summary with improvement recommendations

4.3. TAFES value analysis

To demonstrate the operational relevance of the *TAFES* framework, we revisit the Nurse Linda use-case introduced earlier. This generative AI application in healthcare illustrates how each *TAFES* principle addresses specific gaps in current responsible AI practice.

Transparency: Nurse Linda's decision-support system must provide clear explanations for its recommendations, especially when advising on patient care. *TAFES* operationalizes this through explanability protocols that ensure clinical staff can interrogate and understand AI-generated outputs.

Accountability: The framework delineates responsibility across stakeholders—developers, hospital administrators, and clinical users—using a RACI-style model. This ensures traceability of decisions and clarifies who is answerable for outcomes, particularly in high-stakes scenarios.

Fairness: *TAFES* mandates demographic representativeness in training data and outcome parity across patient groups. In Nurse Linda's case, this prevents biased recommendations that could disadvantage certain populations, such as elderly or minority patients.

Ethics: Ethical deliberation is embedded throughout the lifecycle. For Nurse Linda, this includes assessing the moral implications of delegating care decisions to AI and ensuring alignment with professional codes of conduct and patient dignity.

Safety: Beyond technical robustness, *TAFES* integrates procedural safeguards such as override mechanisms and continuous monitoring. These ensure that Nurse Linda's system minimizes harm and remains within its intended scope of operation.

This value analysis affirms that *TAFES* is not merely a conceptual model but a practical guide for lifecycle governance. It bridges ethical intent with engineering execution, enabling responsible deployment of AI in sensitive domains.

4.4. Challenges and considerations for AI DevOps

The agile approach and rapid prototyping using vibe coding are increasingly prevalent in the implementation of GenAI in many sectors. This presents unique challenges that require careful consideration of *TAFES* principles throughout the DevOps lifecycle. The enhanced implementation of *TAFES* principles in the Nurse Linda AI chatbot presents several technical and organizational challenges, as identified in recent literature on AI ethics and governance (Corrêa NK *et al.* [31], Prunkl *et al.* [25], Qumer *et al.* [24], Raquib *et al.* [28], Sadek *et al.* [27]).

(a) Technical Implementation Challenges:

Explanability Complexity: Balancing system transparency with the complexity of LLM decision-making in healthcare contexts requires sophisticated explanation generation that maintains clinical accuracy while ensuring patient comprehension (Taylor [53]).

Real-time Bias Detection: Implementing continuous bias monitoring across multiple demographic dimensions while maintaining system performance requires optimized algorithms and substantial computational resources (Barocas *et al.* [54]).

Safety Monitoring Integration: Combining clinical safety requirements with AI safety protocols requires careful integration of medical guidelines, regulatory compliance, and technical safety measures (Corrêa NK *et al.* [31]).

(b) Organizational Challenges:

Stakeholder Training: Ensuring all healthcare staff understand AI capabilities, limitations, and *TAFES* implementation requires comprehensive training programs and ongoing education (Sadek *et al.* [27]).

Responsibility Attribution: Establishing clear accountability in AI-assisted healthcare requires careful definition of roles between AI systems, healthcare providers, and institutional oversight (Prem *et al.* [55]).

Cultural Adaptation: Implementing ethics and fairness considerations across diverse patient populations requires cultural sensitivity and adaptation mechanisms (Khan [29], Raquib *et al.* [28]).

(c) Measurement and Validation Challenges:

How might we measure the effectiveness of the enhanced Nurse Linda AI chatbot in terms of *TAFES* compliance? By adhering to the enhanced *TAFES* principles, we could validate whether the clinical support system demonstrates measurable improvements across multiple dimensions:

Enhanced Transparency: Multi-layer explanability with 95% patient comprehension rates and 90% clinician satisfaction with explanation quality.

Comprehensive Accountability: Complete audit trail coverage with 100% incident traceability and sub-2-minute escalation response times.

Measurable Fairness: Bias metrics within acceptable thresholds (< 5% demographic parity violation) across all patient populations with automated correction procedures.

Validated Ethics: 100% compliance with medical ethics guidelines and 95% patient satisfaction with consent and autonomy preservation.

Assured Safety: Zero safety incidents with < 1% false positive rate and 99.9% system availability with backup and emergency response capabilities.

(d) Some Aspirational Performance Metrics for Validation

Transparency Metrics:	Explanation completeness score (target: > 90%)	User comprehension rate (target: > 85% across all demographic groups)	Explanation request fulfillment time (target: < 2 seconds)
Accountability Metrics:	Audit trail completeness (target: 100%)	Incident response time (target: < 2 minutes)	Responsibility clarity score from stakeholder surveys (target: > 90%)
Fairness Metrics:	Demographic parity difference (target: < 5%)	Equalized odds difference (target: < 5%)	Treatment recommendation consistency across groups (target: > 95%)
Ethics Metrics:	Ethics guideline compliance rate (target: 100%)	Patient autonomy preservation score (target: > 95%)	Cultural sensitivity rating from diverse patient groups (target: > 90%)
Safety Metrics:	System availability (target: > 99.9%)	Security incident rate (target: 0 per month)	Patient safety incident rate (target: 0 per quarter)

We further propose a variant of the well-known Turing Test to validate a comprehensive RAI implementation. A panel of domain experts evaluates the system's responses, explanations, and *TAFES* principle adherence across representative scenarios. The panel assesses whether AI recommendations, explanations, and ethical reasoning are distinguishable from recommendations from human experts and medical guidelines. If the evaluation panel cannot distinguish between AI and human performance on TAFES criteria, the system (in this case Nurse Linda, the AI Agent) may claim that it demonstrates "responsible" operation and continued deployment is viable.

5. Limitations and future work

5.1. Reflections on TAFES using the lens of critical realism

While the *TAFES* framework provides comprehensive guidance for responsible AI implementation, several deeper challenges and philosophical tensions warrant discussion. Using the lens of Critical Realism, a research philosophy that seeks to explain social phenomena by identifying the unobservable, underlying "generative mechanisms" that cause observable events, we may distil the following mechanisms that lead to implementation outcomes.

Conflicting Principles: Certain principles may conflict in practice, most notably Transparency and Privacy. Complete algorithmic transparency can inadvertently expose sensitive personal or proprietary data. *TAFES* mitigates this by promoting proportional transparency, and providing sufficient explanability to ensure trust and oversight without breaching privacy or security.

Applicability in Low-Resource Settings: Implementing *TAFES* in small enterprises or Global South contexts may be constrained by limited resources and expertise. The framework encourages scalable adoption, where baseline compliance emphasizes Fairness, Accountability, and Safety as a minimum viable standard, with progressive adoption of Transparency and Ethics as capacity grows. Jurisdictional Variability: Regulatory expectations differ significantly across regions. The EU's AI Act employs strict compliance models, whereas other jurisdictions adopt principle-based or voluntary approaches. *TAFES* accommodates these differences through modular design, allowing contextual interpretation while retaining coherence.

Philosophical Coherence: Critics might view the combination of utilitarian safety principles, deontological accountability, and virtue-based ethics as inconsistent. However, this plurality reflects real-world governance, where multiple ethical traditions coexist to balance human welfare, duty, and integrity. The pluralist stance strengthens, rather than weakens, *TAFES* by aligning diverse normative expectations into a live, socio-technical system.

These reflections reinforce *TAFES*'s capacity for critical self-assessment and evolution. By remaining reflexive and context-sensitive, the framework can respond dynamically to technological advances and societal expectations while maintaining ethical and procedural rigour.

Beyond the philosophical discussion, the technical challenges confronting the DevOps environment remain far too complex. The interlock of law, ethics, and technology, such as AI, is evolving and has by no means reached common ground for the implication of legal liability. This confounds the interpretation of whether an adequate duty of care was demonstrated by the various parties involved in the development and operation of such systems. Alarmingly, when the proverbial genies are out of the bottle, it is not going to be easy to put them back! Lyons [56] reported that Google DeepMind, for instance, had identified a new AI threat scenario to its AI Safety manual. "A model might try to prevent its operators from modifying it or shutting it down." A panel of 200, including 10 Nobel Laureates, have signed a letter calling on the United Nations to draw a red line. It reads, "Some advanced AI systems have already exhibited deceptive and harmful behavior, and yet these systems are being given more autonomy to take actions and make decisions in the world," … AI "could soon far surpass human capabilities and escalate risks such as engineered pandemics, widespread disinformation, large-scale manipulation of individuals including children, national and international security concerns, mass unemployment, and systematic human rights violations (Thomson [57])."

In the face of such grave risks and uncertainties, we acknowledge the following challenges in moving forward.

Implementation Complexity: The comprehensive nature of *TAFES* implementation requires significant technical expertise, organizational resources, and sustained commitment that may exceed the capabilities of smaller organizations or resource-constrained environments (Sadek *et al.* [27]).

Cultural and Contextual Adaptation: The framework's effectiveness across diverse cultural contexts and regulatory environments requires adaptation that we have not fully validated through empirical testing across global implementations (Corrêa NK *et al.* [31]).

Measurement and Quantification Challenges: While technical metrics for bias and safety can be quantified objectively, broader ethical and fairness considerations often involve subjective assessments that vary across stakeholders and contexts (Prem *et al.* [55]).

Technological Evolution Pace: The rapid advancement of AI technology, particularly in generative AI and foundation models, may outpace framework development and require continuous evolution to address emerging capabilities and risks [52].

Hence, we may concede that there are empirical validation gaps that must be addressed prior to field application of the *TAFES* framework.

5.2. Moving from reflection to action

Sharma *et al.* [14] rhetorically quoted the Editor in Chief of IEEE Computer, "Is our question of zero-trust AI even worthy of discussion when AI is already ubiquitous and embedded in almost everything we rely on? So, maybe AI is simply a new, hidden, and unavoidable risk to life, devoid of opt-out options." We have since concluded that the perceived or experienced virtues of AI systems cannot be an afterthought and require community-driven solutions. The responsible use of AI in governance, health care, education, and industry remain impeded by common pitfalls.

Limited Longitudinal Evidence: Current validation relies primarily on theoretical analysis and single-point assessments rather than longitudinal studies demonstrating sustained effectiveness over extended deployment periods.

Stakeholder Engagement Breadth: While expert consultation informed framework development, comprehensive engagement across diverse communities, particularly marginalized groups most affected by AI bias, requires expansion.

Cross-Sector Validation: The healthcare case study provides sector-specific validation, but comprehensive testing across domains (finance, education, criminal justice) remains limited.

In an attempt to address these limitations, we propose the following fruitful avenues for further research:

(1) Practical Implementation Guidelines: As called for in the NIST strategy [52], regulators, in collaboration with global counterparts, industry, academia, and consumer groups, should develop practical guidelines for integrating *TAFES* considerations across the AI system lifecycle with sector-specific adaptations.

(2) Cultural and Regional Impact: Investigate how cultural and regional differences influence the implementation of RAI principles through diverse field studies of use-cases, particularly incorporating Global South perspectives on technological sovereignty and collective benefit approaches.

(3) Adaptable *TAFES* Standards: Despite differences in socio-economic context, comparative analysis of existing AI regulatory frameworks reveals opportunities for developing more comprehensive and unified procedures and methods for establishing compliance across diverse regulatory environments.

(4) Longitudinal Studies: Track the real-world implementation of the *TAFES* framework in the design, development, and deployment of AI systems across sectors to assess effectiveness, challenges, and refinement opportunities over extended periods.

(5) Monitoring and Control Dashboard: Creating and validating customizable dashboards for monitoring and controlling *TAFES* measures throughout the AI lifecycle, building upon initiatives from technology companies such as Microsoft [58], IBM [59] and many others [47]. This involves stakeholders in identifying, measuring, and evaluating essential performance metrics of AI systems while setting thresholds for acceptable (continued deployment) or unacceptable (requiring decommissioning) outcomes.

While the *TAFES* framework offers a comprehensive and implementable model for responsible AI, the above challenges warrant ongoing reflection, discussion, and debate among scholars, developers, and regulators. These reflections highlight the importance of ongoing dialogue, empirical validation, and iterative refinement. *TAFES* provides a principled foundation, but its success hinges on collaborative engagement and a sustained commitment to ethical AI governance.

6. A provisional manifesto for RAI

The *TAFES* Framework for RAI represents a pivotal step toward assuring the responsible design, development, deployment, and decommissioning of AI systems. Rooted in techno-moral virtues (Sharma *et al.* [14]) and the aspiration of contextual ethics (Sharma *et al.* [2]), this manifesto seeks to maximize positive outcomes while mitigating negative affordances. In this concluding section, we present the essence of a provisional RAI manifesto and examine its implications for practice.

(1) Balancing Affordances and Decommissioning: The *TAFES* approach recognizes that AI systems offer both positive and negative affordances. A critical juncture arises when negative consequences outweigh the benefits or when the positive impact fails to justify the risks. By integrating these considerations into the AI-based systems life cycle, we strive for alignment with *TAFES* principles, fostering responsible outcomes for society.

(2) Dialogues Across Stakeholders: To achieve RAI, collaboration among stakeholders is paramount. The design, development, and deployment of AI-based systems should commence with transparent dialogues. These conversations involve technology service providers, government regulators, and user communities. Academic scholars, as unbiased interlocutors, play a crucial role in shaping these discussions.

(3) Seeking A Common Understanding: Our proposed Manifesto for RAI hinges on a shared understanding among stakeholders. Academics, user groups, regulators, and technology providers must continually address any techno-moral challenges that emerge. This involves setting safeguards, monitoring compliance, and promoting openness in platforms, algorithms, and explanations. As we have observed, the current conundrum with generally accepted AI safety standards is that none exist with universal adoption. And this is because the current outreach towards ethics, safety and responsibility from technology firms and GenAI-enabled service-providers seem to be an afterthought.

(4) Techno-Moral Virtues and Ethics: Applying *TAFES* principles must result in non-functional, yet essential characteristics such as the *TAFES* affordances [16].

(5) Advocacy for Responsible Practices: More specifically, we advocate for specific actions across key domains:

- Researchers & Scholars must develop operational metrics for AI trustworthiness. Specifically, create real-time monitoring dashboards that track hallucination rates across LLM outputs, quantify bias through fairness metrics (e.g., demographic parity, equalized odds), and establish SLA threshold. For instance, requiring < 5% hallucination rates or explainability scores above defined benchmarks before clinical deployment.
- Regulators must mandate algorithmic audits with enforcement mechanisms. Beyond frameworks like the EU AI Act, establish inspection protocols that require: public disclosure of training data sources for high-risk applications, quarterly bias audits with published results, and penalty structures (e.g., 2%–4% global revenue fines) for non-compliance; mirroring GDPR's enforcement model.
- Big Tech & AI Service Providers must implement TAFES-by-design principles with measurable commitments. Concrete actions include: publishing model cards detailing training data composition and known limitations, establishing independent ethics boards with veto power over deployments [47], and adopting open-source models for public sector applications where transparency requirements are paramount.
- Users & Consumers must gain AI literacy through accessible education and participatory rights. Deploy mandatory plain-language disclosures when AI influences decisions (healthcare diagnoses, loan approvals), establish citizen review panels for high-impact AI systems, and create simple opt-out mechanisms, thus ensuring users can meaningfully exercise agency over AI-mediated interactions.

(6) Building compliance and conformance toolkits that align with international standards: The *TAFES* framework aligns with prior understanding of safe and RAI frameworks [3,4,33,36,39,42,42,48]). These frameworks require AI development teams to map, measure, and manage risks for AI applications throughout their development lifecycle. To the extent feasible, the degree of commitment to *TAFES* attributes should be specified at the onset of the AI lifecycle, so that concerned stakeholders are aware and can act accordingly.

In short, our Manifesto for RAI is a clarion call for a holistic approach; one that balances technological progress with ethical imperatives. By adhering to *TAFES* principles, we may work hope that AI technologies benefit society while upholding fundamental values. Our contemplation extends to how society can optimally harness AI's potential. We question whether diverse ethical perspectives align with universal design principles for responsible, trustworthy, and techno-moral AI. The intended impact of the *TAFES* principles is to articulate a comprehensive manifesto for RAI implementation. But they are not without discord. As examples:

First, principle conflicts may arise. For example, the imperative for Transparency can conflict with Privacy, especially in domains like healthcare or finance. *TAFES* acknowledges these tensions and encourages context-sensitive trade-offs guided by stakeholder deliberation and ethical reasoning.

Second, jurisdictional and cultural variability complicates implementation. Legal definitions of fairness or accountability differ across regions, and ethical norms are not universally shared. *TAFES* addresses this by promoting contextual adaptability, allowing regulators and organizations to calibrate emphasis on specific principles without compromising core integrity.

Third, resource constraints in low-income settings may hinder full lifecycle implementation. While *TAFES* prescribes robust governance mechanisms, it also supports scalable adoption by prioritizing principles like Fairness and Safety in constrained environments. Future work will explore lightweight adaptations of *TAFES* for such contexts.

Fourth, stakeholder accountability remains a contested issue. The question of who is ultimately responsible for AI outcomes—users, developers, or regulators—requires clearer delineation. *TAFES* contributes to this discourse by embedding structured accountability mechanisms and promoting shared responsibility across the AI ecosystem.

There is an urgency for regulating AI, especially with respect to "interpretability, *i.e.*, in understanding the inner workings of AI systems, before models reach an overwhelming level of power" (Amodei [60]). Given the prevalence of GenAI, including ChatGPT and similar tools embedded in widely used software, such as search engines, content authoring platforms, and data analytics tools a dialogue on RAI becomes not only feasible but imperative. Jobin *et al.* [1] reveal that national and international organizations, from Australia to the United Kingdom, have responded to ethical AI concerns by establishing expert committees mandated to draft policy documents. However, consensus remains elusive. Even within the EU, despite the AI Act being implemented in 2024, challenges remain in enforcement and universal adoption across member states. While such a multifaceted ethical landscape remains a challenge, the prior work reported in this article gives cause for a belief that AI must serve a universal good. These aspirations resonate globally and warrant attention, even from international bodies like the United Nations. Yet, the challenge lies in translating these lofty ideals into practical implementation.

While the *TAFES* framework provides a foundational structure, several limitations require acknowledgment and future research attention. In the Design Phase, there is need for more specific methodologies and tools for embedding ethical considerations during AI system design. In the Development Phase, neither existing IEEE nor ISACA guidelines provide sufficient professional expertise in translating high-level *TAFES* principles into actionable development steps. In the Deployment Phase, maintaining safety and responsibility standards throughout the system's lifecycle requires continuous evolution of monitoring and oversight mechanisms. And finally, in the Decommissioning Phase, insufficient guidance exists for identifying, measuring, and controlling non-compliance to *TAFES* that necessitates system shutdown and responsible data disposal.

To conclude, our journey toward RAI requires collective effort, interdisciplinary collaboration, and ongoing refinement. As the Microsoft [58] responsible AI Transparency Report declares, "Technology sector-led initiatives comprise one important force to advance RAI. Industry and others stand to significantly benefit from the key role that governments can also play." We conjecture that such consensus-based safety frameworks may prove more applicable, usable and enforceable when they serve as the starting point for developing standards and guidelines, rather than as an afterthought. We intend the *TAFES* framework to serve as the basis for further thought and discussion, guiding us toward a future where AI benefits humanity while upholding universal values. The imperative for regulating AI has reached a critical juncture globally. Failure to promote RAI will invariably result in public skepticism and resistance to AI adoption, potentially hindering the development of beneficial applications. The *TAFES* framework provides a robust foundation for RAI development and deployment. As outlined in our future research directions, continued work will focus on developing concrete methodologies to seamlessly embed ethical considerations throughout the AI system life cycle, thereby ensuring trust and

confidence in its societal impact. This represents an affirmation of the significant conversations taking place around regulatory and governance issues for AI, a fraction of which we have referenced in this paper.

Acknowledgements

The authors are grateful to the editor and anonymous reviewers for their thoughtful comments and constructive feedback. Many thanks are also due to numerous colleagues who took the time to debate and discuss many of the ideas expressed in this paper. An early version of this paper was presented at a seminar at Santa Clara University, USA.

Authors' contribution

Conceptualization, SL, RS, NK and AZ; methodology, SL; software, AZ; validation, SL, RS, NK and AZ; formal analysis, SL and RS; investigation, SL, RS, NK and AZ; resources, RS; data curation, SL; writing—original draft preparation, SL, RS, AZ; writing—review and editing, NK, AZ, RS; project administration, SL; funding acquisition, RS. All authors have read and agreed to the published version of the manuscript.

Conflicts of interests

The authors declare no conflict of interest.

References

[1] Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* 2019, 1(9):389–399.

[2] Sharma RS, Loucif S, Kshetri N, Voas J. Global initiatives on "safer" and more "responsible" artificial intelligence. *Computer* 2024, 57(11):131–137.

[3] IBM. Responsible AI. 2022. Available: https://www.ibm.com/trust/responsible-ai (accessed on 14 September 2025).

[4] Microsoft. Microsoft responsible AI Standard Reference Guide. 2022. Available: https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/bade/documents/products-and-services/en-us/ai/RAIS-Reference-Guide-v2.pdf (accessed on 14 September 2025).

[5] Microsoft. Governing AI—a blueprint for the future. 2023. Available: https://www.linkedin.com/pulse/governing-ai-blueprint-future-microsoft-on-the-issues/ (accessed on 14 September 2025).

[6] Google. AI Principles. 2022. Available: https://ai.google/principles/ (accessed on 14 September 2025).

[7] Google AI. Perspectives on Issues in AI Governance. 2022. Available: https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf (accessed on 14 September 2025)

[8] Gong J, Qu H, Dorwart H. AI governance in China: strategies, initiatives, and key considerations. 2024. Available: https://www.twobirds.com/en/insights/2024/china/ai-governance-in-china-strategies-initiatives-and-key-considerations (accessed on 14 September 2025).

[9] European Commission. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). 2024. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng (accessed on 14 September 2025).

[10] European Commission. AI Act. 2024. Available: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (accessed on 14 September 2025).

[11] G20. New Delhi Leaders' Declaration. 2023. Available: https://www.mea.gov.in/Images/CPV/G20-New-Delhi-Leaders-Declaration.pdf (accessed on 14 September 2025).

[12] Floridi L. The 7 Good AI Global Frameworks. 2020. Available: https://www.eismd.eu/wp-content/uploads/2019/12/AI4People-2020.pdf (accessed on 14 September 2025).

[13] Polat G. Unlocking AI's potential: How COBIT Can Guide Your Business Transformation. 2024. Available: https://www.isaca.org/resources/news-and-trends/isaca-now-blog/2024/unlocking-ais-potential-how-cobit-can-guide-your-business-transformation (accessed on 14 September 2025).

[14] Sharma R, Ahmad N, Ali S, Bilal A, Fatima S, *et al*. Techno moral affordances of artificial intelligence in data-driven systems. *Computer* 2020, 55(10):76–81.

[15] Sharma RS, Loucif S, Khalil A, Zahid A. A manifesto for responsible AI: healthcare use-case of the TAFES framework. In *Proceedings of Eighth International Conference on Information System Design and Intelligent Applications*, Dubai, United Arab Emirates, January 3–4, 2025.

[16] Bilal A, Wingreen S, Sharma R, Jahanbin P. Trust development in artificial intelligence-based emerging technologies: rise of techno moral virtues and data ethics. In *Australasian Conference on Information Systems 2021 (ACIS 2021)*, Sydney, Australia, December 6–10, 2021.

[17] König PD, Wurster S, Siewert MB. Consumers are willing to pay a price for explainable, but not for green AI. Evidence from a choice-based conjoint analysis. *Big Data Soc.* 2022, 9(1):20539517211069632.

[18] Dignum V. The myth of complete AI-fairness. In *Artificial Intelligence in Medicine: Proceedings of the 19th International Conference on Artificial Intelligence in Medicine, AIME 2021*, Online, June 15–18, 2021, pp. 3–8.

[19] Dignum V. Responsible AI: from principles to action. In *Robophilosophy 2022: Social Robots in Social Institutions*, Helsinki, Finland, August 16–19, 2020, p. 13.

[20] Floridi L. Translating principles into practices of digital ethics: five risks of being unethical. *Philos. Technol.* 2019, 32:185–193.

[21] Floridi L, Cowls J. A unified framework of five principles for AI in society. *Harvard Data Sci. Rev.* 2019, 1(1):1–15.

[22] Floridi L, Cowls J, King TC, Taddeo M. How to design AI for social good: seven essential factors. *Sci. Eng. Ethics* 2020, 26(3):1771–1796.

[23] Das S, Green BP, Varshney K, Ganapini M, Renda A. Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society (AIES 2024). *ACM FAccT Conf. Ser.* 2024.

[24] Qumer SM. Timnit Gebru: seeking to promote diversity and ethics in AI. 2023. Available: https://doi.org/10.1108/CFW-07-2022-0019 (accessed on 14 September 2025).

[25] Prunkl CEA, Ashurst C, Anderljung M, Webb H, Leike J, *et al*. Institutionalizing ethics in AI through broader impact requirements. *Nat. Mach. Intell.* 2021, 3(2):104–110.

[26] Sinha S, Lee YM. Challenges with developing and deploying AI models and applications in industrial systems. *Discover Artif. Intell.* 2024, 4(1):55.

[27] Sadek M, Kallina E, Bohné T, Mougenot C, Calvo RA, *et al*. Challenges of responsible AI in practice: scoping review and recommended actions. *AI Soc.* 2025, 40(1):199–215.

[28] Raquib A, Channa B, Zubair T, Qadir J. Islamic virtue-based ethics for artificial intelligence. *Discover Artif. Intell.* 2022, 2:11.

[29] Khan G. Islamic AI. 2024. Available: https://www.amazon.sg/Islamic-AI-Gohar-Khan/dp/B0CZMSHK1P#detailBullets_feature_div (accessed on 14 September 2025).

[30] Cheng J, Zeng J. Shaping AI's future? China in global AI governance. *J. Contemp. China* 2022, 32(143):794–810.

[31] Corrêa NK, Galvão C, Santos JW, Del Pino C, Pinto EP, *et al*. (2024). Worldwide AI ethics: a review of 200 guidelines and recommendations for AI governance. *Patterns* 2023, 4(10):100857.

[32] Kinney D. Aggregating concepts of fairness and accuracy in prediction algorithms. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*, Athens, Greece, June 23–26, 2025, pp. 464–472.

[33] European Commission. Ethics guidelines for trustworthy AI. 2019. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed on 14 September 2025).

[34] FLIA. Artificial Intelligence Development Plan. 2017. Available: https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm (accessed on 14 September 2025).

[35] Khanal S, Zhang H, Taeihagh A. Development of new generation of Artificial Intelligence in China: when Beijing's global ambitions meet local realities. *J. Contemp. China* 2025, 34(151):19–42.

[36] NITI Aayog. National strategy for artificial intelligence. 2018. Available: https://www.niti.gov.in/writereaddata/files/AI-Strategy-Document.pdf (accessed on 14 September 2025).

[37] IMPR. IndiaAI Mission (2024): Empowering Innovation, Infrastructure, & Inclusive Growth through AI. 2025. Available: https://www.impriindia.com/insights/indiaai-mission-2024/ (accessed on 14 September 2025).

[38] Brazilian Government (Ministry of Science, Technology and Innovation). Brazilian Artificial Intelligence Plan 2024–2028. 2024. Available: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/plano-brasileiro-de-inteligencia-artificial (accessed on 14 September 2025).

[39] South African Department of Communications & Digital Technologies. Artificial Intelligence Policy Framework. 2024. Available: https://www.dcdt.gov.za/sa-national-ai-policy-framework/file/338-sa-national-ai-policy-framework.html (accessed on 14 September 2025).

[40] Artificial Intelligence Office UAE. UAE AI Ethics Principles & Guidelines Dec 2022. 2024. Available: https://ai.gov.ae/wp-content/uploads/2023/03/MOCAI-AI-Ethics-EN-1.pdf (accessed on 14 September 2025).

[41] PDPC. Singapore's Approach to AI Governance. 2020. Available: https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework (accessed on 14 September 2025).

[42] Singapore IMDA. Model AI Governance Framework 2024. Available: https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2024/public-consult-model-ai-governance-framework-genai (accessed on 14 September 2025).

[43] DISR Australia. Voluntary AI Safety Standard. 2024. Available: https://www.industry.gov.au/publications/voluntary-ai-safety-standard (accessed on 14 September 2025).
[44] U.S. Department of Commerce (National Institute of Standards and Technology). NIST Trustworthy and responsible AI (NIST AI 100-5): A Plan for Global Engagement on AI Standards. 2024. Available: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-5.pdf (accessed on 14 September 2025).
[45] UK Government (Department for Science, Innovation and Technology; Office for Artificial Intelligence). A pro-innovation approach to AI regulation: policy paper. 2023. Available: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper (accessed on 14 September 2025).
[46] UK Government. A pro-innovation approach to AI regulation. 2023. Available: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper (accessed on 14 September 2025).
[47] Lohchab H. AI ethics boards coming up fast in Indian tech majors. 2024. Available: https://economictimes.indiatimes.com/tech/artificial-intelligence/ai-ethics-boards-coming-up-fast-in-indian-tech-majors/articleshow/112814960.cms?from=mdr (accessed on 14 September 2025).
[48] IEEE SA. The IEEE Global Initiative 2.0 on Ethics of Autonomous and Intelligent Systems. Available: https://standards.ieee.org/industry-connections/activities/ieee-global-initiative/ (accessed on 14 September 2025).
[49] Future of Life Institute. Asilomar AI Principles. 2017. Available: https://futureoflife.org/ai-principles (accessed on 14 September 2025).
[50] ISACA. Considerations for Implementing a Generative Artificial Intelligence Policy. 2024. Available: https://www.isaca.org/resources/ebook/considerations-for-implementing-a-generative-artificial-intelligence-policy (accessed on 14 September 2025).
[51] Bughin J. Doing versus saying: responsible AI among large firms. *AI Soc.* 2024, 40(4):2751–2763.
[52] National Institute of Standards and Technology (NIST). A Plan for Global Engagement on AI Standards. 2024. Available: https://doi.org/10.6028/NIST.AI.100-5 (accessed on 14 September 2025).
[53] Taylor I. Is explainable AI responsible AI? *AI Soc.* 2025, 40:1695–1704.
[54] Barocas S, Hardt M, Narayanan A. *Fairness and Machine Learning: Limitations and Opportunities*, 1st ed. Cambridge: MIT Press, 2023.
[55] Prem E. From ethical AI frameworks to tools: a review of approaches. *AI Ethics* 2023, 3(4):699–716.
[56] Lyons J. AI gone rogue: models may try to stop people from shutting them down, Google warns. 2025. Available: https://www.theregister.com/2025/09/22/google_ai_misalignment_risk/ (accessed on 14 September 2025).
[57] Thomson I. Stop runaway AI before it's too late, experts beg the UN. 2025. Available: https://www.theregister.com/2025/09/23/ai_un_controls/ (accessed on 14 September 2025).
[58] Microsoft. Responsible AI Transparency Report. 2025. Available: https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/Responsible-AI-Transparency-Report-2025-vertical.pdf (accessed on 14 September 2025).
[59] IBM. A look into IBM's AI ethics governance framework. 2024. Available: https://www.ibm.com/think/insights/a-look-into-ibms-ai-ethics-governance-framework (accessed on 14 September 2025).
[60] Amodei D. The Urgency of Interpretability. 2025. Available: https://www.darioamodei.com/post/the-urgency-of-interpretability (accessed on 14 September 2025).