Journal of International DBA Studies (JIDS)
Vol. 1 No. 001 (2025) https://eaj.ebujournals.lu/index.php/JIDS/index

RAISEF: A Driver-Based Framework for Responsible AI Integrating Academic and Practical Perspectives

by
Richard R. Khan*

Abstract

AI is rapidly proliferating in many industries, presenting unprecedented opportunities and challenges impacting humanity across ethical, regulatory, and societal dimensions. This study introduces the Responsible AI System Evolution Framework or RAISEF. It is a novel lifecycle-based approach that helps stakeholders navigate evolving challenges, bridging theoretical constructs with practical applications. It organizes 15 interdependent, Responsible AI drivers into three pillars: ethical safeguards, operational integrity, and societal empowerment. RAISEF systematically addresses many inter-driver tensions such as privacy versus explainability and fairness versus robustness. Additionally, it promotes synergies for cohesive, Responsible AI implementation. Furthermore, unlike existing models, RAISEF integrates cross-disciplinary insights from ethics, governance, sociology, and systems thinking. It advances the theoretical discourse while offering actionable methodologies and toolkits tailored to diverse cultural and environmental contexts. The paper, through mainly hypothetical and empirical scenarios, illustrates RAISEF's adaptability to emerging challenges, including autonomous systems, generative AI, and global policy variations. It unites theory and practice. RAISEF provides a comprehensive, globally adaptable framework for academics, policymakers, and practitioners to foster the development of ethical, sustainable, and trustworthy AI systems.

Keywords: Artificial intelligence, Responsible AI, ethical safeguards, operational integrity, societal empowerment

* Richard R. Khan is an AI/ML and IT executive with more than 20 years of experience in digital transformation and software engineering, and currently the Vice President of Software Development, driving enterprise AI strategies that cut costs and grow revenues.

Introduction

Artificial intelligence now permeates business, government, and daily life, progressing from rule-based systems to data-driven and generative models that promise unprecedented efficiency and innovation (Russell & Norvig, 2021). High-profile privacy breaches, opacity, and biased outputs, however, reveal profound ethical and societal risks (Floridi et al., 2018; Jobin et al., 2019). Rising public demands for fairness, accountability, and transparency (FAT) press developers to embed protective design principles and regulators to enact safeguards, as seen in the EU AI Act (Office of the European Union, 2024).

Responsible AI initiatives still lack an integrated lens that carries accountability and trustworthiness through the entire development lifecycle (Jobin et al., 2019). Existing schemes often isolate tasks, such as bias mitigation or regulatory compliance, without treating their interdependence (Floridi et al., 2018).

The Responsible AI System Evolution Framework (RAISEF) closes this gap by interweaving fifteen mutually reinforcing drivers, organised into ethical safeguards, operational integrity, and societal empowerment, across every lifecycle phase. Sector-specific and culturally sensitive guidance turns these principles into actionable practice for executives and project sponsors. RAISEF fuses sociological, economic, and environmental perspectives. Combining fairness-aware algorithms with social science DEI principles yields tools tailored to marginalized or resource-constrained contexts. By tracking how its fifteen drivers interact from development through deployment and oversight, the framework surfaces tensions, such as bias mitigation versus explainability, highlighting the need for an integrated, value-aligned methodology (Busuioc, 2021; Rudin, 2019).

Responsible AI must mature as a sociotechnical enterprise that embeds societal, legal, and moral norms, replacing performance-only targets with value-driven outcomes secured by governance, transparency, and accountability (Bullock et al., 2024; Dubber et al., 2020). This paper follows that logic: the introduction states objectives and unveils RAISEF; the literature survey outlines its fifteen drivers; the discussion analyses their interplay; the conclusion maps implications and future research.

RAISEF supplies a springboard for further study. Forthcoming doctoral work will empirically refine it. Targeted at experts yet accessible through toolkits and sectoral case studies, a practitioner-oriented implementation guide will follow.

Introducing RAISEF: A Novel Lifecycle Approach

The Responsible AI System Evolution Framework represents a novel approach to Responsible AI, aligning 15 key drivers across all stages, from design to post-deployment

monitoring, within a product and project lifecycle framework. Unlike prior models that focus on organizational maturity or specific phases, RAISEF ensures the comprehensive integration of principles across industries and cultural contexts. It allows the implementation to be tailored to sector-specific challenges using customized toolkits.

The framework categorizes the 15 drivers into three overarching pillars. Ethical safeguards (FIBMAP: fairness, inclusiveness, bias mitigation, accountability, privacy) drivers prioritize protecting human rights, sensitive data, and ethical and moral standards. Operational integrity (GRIESS: governance, robustness, interpretability, explainability, security, safety) drivers focus on ensuring technical reliability, robustness, and compliance with best practices. Societal empowerment (SHOTT: sustainability, human oversight, transparency, trustworthiness) drivers foster public trust and confidence by aligning AI systems with societal values (see Table 1). Figure 1 visualizes the critical drivers for addressing ethical, operational, and societal challenges.

Table 1

Categorization of Responsible AI drivers into ethical safeguards, operational integrity, and societal empowerment pillars

Pillar	Drivers
Ethical Safeguards	Fairness, Inclusiveness, Bias Mitigation, Accountability, Privacy (FIBMAP)
Operational Integrity	Governance, Robustness, Interpretability, Explainability, Security, Safety (GRIESS)
Societal Empowerment	Sustainability, Human Oversight, Transparency, Trustworthiness (SHOTT)

Note. Table created by the author

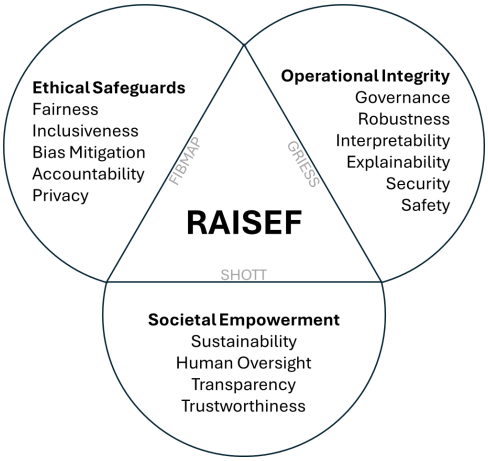
More importantly, RAISEF also addresses the inherent tensions frequently arising during AI system development and deployment between these drivers. For instance, by increasing fairness, there can be a negative impact on operational integrity (specifically, robustness). Each of these trade-offs demands dynamic mechanisms for identifying, analyzing, and mitigating conflicts across the system's lifecycle. Enhancing privacy, to illustrate another example, can also potentially degrade operational integrity (but this time, explainability). RAISEF ensures Responsible AI principles are cohesively implemented and adapted to real-world complexities.

Beyond its theoretical alignment of drivers, RAISEF's emphasis on a product and project lifecycle approach extends its utility. It provides actionable guidance for identifying leverage points where identified tensions can be effectively balanced. The key challenge is maintaining interdependence among the pillars rather than allowing them to compete. It offers a comprehensive roadmap that weaves together theoretical principles with

practical, real-world implementation. This adaptability makes RAISEF globally relevant, addressing cultural inclusivity and sector-specific regulatory demands.

Figure 1

RAISEF Three Pillars



Note. Figure created by the author

Implementing RAISEF Across the AI Lifecycle

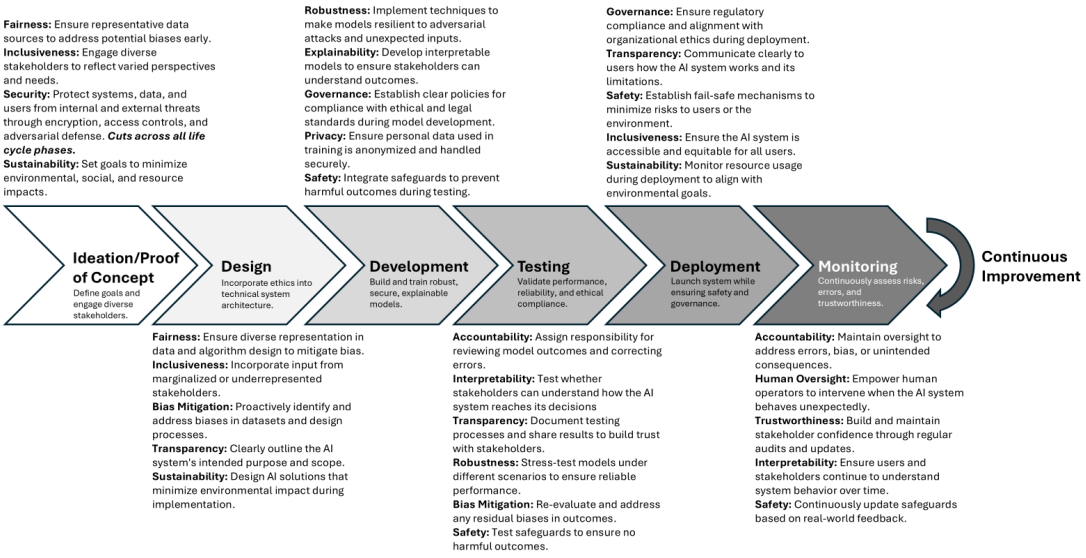
RAISEF systematically integrates 15 Responsible AI drivers that operate across the entire AI lifecycle (see Figure 2). Each lifecycle stage leverages specific drivers to ensure the AI system aligns with ethical, operational, and societal values while addressing interdependencies and trade-offs. By addressing critical issues early in the lifecycle, such as bias mitigation, fairness, and privacy during the ideation and design phases, RAISEF prevents the compounding of problems in later stages. The decisions made in earlier stages of the lifecycle are consistently carried forward and cohesively integrated.

The iterative cycle incorporating inclusiveness in stakeholder engagement or embedding explainability during model development becomes a continuous improvement loop. The continuous cyclical approach minimizes the likelihood of needing to rework or abandon implementations when ethical or technical challenges arise in subsequent and later stages. The approach also establishes trust, accountability, and resilience in the system's design. The embedded mechanisms anticipate tensions, such as those between privacy and transparency, and fairness and robustness, providing pathways that enable a balanced resolution. RAISEF enables the development of AI systems that are robust, sustainable, and aligned with societal values throughout their entire lifecycle.

Ideation/Proof of Concept: Fairness, inclusiveness, and sustainability dominate the earliest stage. Broad stakeholder consultation surfaces marginalised perspectives, and explicit sustainability targets stop social or environmental harms from being "baked in".

Figure 2

AI lifecycle stages aligned with RAISEF



Note. Figure created by the author

Design: Bias mitigation and transparency join the core trio. Architecture reviews and data curation avert discriminatory outputs, while transparent design artefacts explain system logic to all stakeholders, embedding ethical and societal values in the blueprint.

Development: Robustness, explainability, governance, privacy, and safety are integrated. Stress-oriented testing bolsters reliability; XAI tools cultivate trust; formal governance structures and privacy-preserving techniques (such as federated learning) ensure compliance; safety guardrails prevent harmful behaviours.

Table 2 illustrates how RAISEF integrates its fifteen drivers throughout the whole product lifecycle, ensuring that ethical, operational, and societal priorities remain aligned from conception to monitoring. Black circles (●) indicate primary engagement phases while gray circles (●) represent supporting or ongoing relevance. Security, for example, begins at ideation and remains critical through deployment and monitoring.

Table 2

RAISEF Driver Engagement Across Life Cycle Phases

	Project/Product Life Cycle Phase					
	Ideation/ Proof of Concept	Design	Development	Testing	Deployment	Monitoring
Ethical Safeguards (FIBMAP)						
Fairness	●	●	●	●	●	●
Inclusiveness	●	●	●	●	●	●
Bias Mitigation		●	●	●	●	●
Accountability				●	●	●
Privacy			●	●	●	●
Operational Integrity (GREISS)						
Governance			●	●	●	●
Robustness			●	●	●	●
Interpretability				●	●	●
Explainability			●	●	●	●
Security	●	●	●	●	●	●
Safety			●	●	●	●
Societal Empowerment (SHOTT)						
Sustainability	●	●	●	●	●	●
Human Oversight						●
Transparency		●	●	●	●	●
Trustworthiness						●

Note. Table created by the author

Testing: Accountability and interpretability are added without losing sight of prior drivers. Audit trails, decision logs, and interpretability checks enable independent verification, while stress tests, safety drills, and renewed bias scans confirm risk boundaries.

Deployment: Governance re-escalates in tandem with transparency, safety, inclusiveness, and sustainability. Live compliance monitoring, accessible interfaces for diverse users, impact assessments, and candid disclosures reinforce responsible roll-out.

Monitoring: Continuous feedback fosters accountability, human oversight, trustworthiness, interpretability, and safety. Stakeholder input triggers updates, and interpretability tools sustain user confidence. Persistent governance and safety protocols guard long-term reliability.

By uniting ethical, operational, and societal priorities at both micro and macro scales, RAISEF provides a globally adaptable path to trustworthy, scalable, and sustainable AI, spanning from clinical decision support to automated credit scoring, and meeting the demands of an increasingly diverse and fast-moving world. This top-down and bottom-up integration allows organizations to elevate or de-emphasize drivers to suit sectoral and cultural contexts, turning discrete projects into a coherent, mutually reinforcing ecosystem. This approach enables context-specific challenges to feed into a consistent organizational strategy.

Literature Review of Responsible AI Drivers

Responsible AI aligns technological progress with human values by reducing bias, ensuring transparency, and adhering to legal norms (Batool et al., 2023; Lu et al., 2024). Core principles, including accountability, explainability, privacy, and inclusive engagement, build trust and equity across the lifecycle (Bullock et al., 2024; Dubber et al., 2020). Governance that integrates ethical, social, and technical dimensions operationalizes the 15 drivers, including fairness, bias mitigation, and trustworthiness, while acknowledging the associated tensions.

RAISEF integrates these drivers via an inter-driver matrix that marks relationships as tensioned, reinforcing, or neutral, enabling agile management of trade-offs among transparency, accountability, and privacy. Inclusiveness mediates fairness and explainability; governance coupled with environmental assessment strengthens sustainability without eroding robustness.

Trustworthiness arises collectively as drivers co-evolve through lifecycle phases and sectors. With 105 mapped interactions, RAISEF enables organizations to reprioritize, for example, by emphasizing inclusiveness during healthcare development and robustness at deployment, thereby tailoring Responsible AI to cultural and regulatory contexts (Bommasani et al., 2021; Bullock et al., 2024). This ecosystem approach replaces checklist compliance with continuous, collaborative adaptation among regulators, industry, and communities.

Ethical Safeguards, Operational Integrity, and Societal Empowerment

RAISEF groups its 15 drivers under three pillars: ethical safeguards that protect human rights; operational integrity that ensures technical reliability; and societal empowerment that aligns AI with community values.

By threading these pillars from ideation to monitoring, RAISEF resolves privacy-explainability and other trade-offs that fragment compliance-driven models (IBM, 2024; Google, 2023; Dubber et al., 2020). Practical toolkits demonstrate, for example, how healthcare diagnostics strike a balance between fairness, accountability, and robustness, while credit scoring integrates bias mitigation with transparency and sustainability. Thus, RAISEF merges theory with action, delivering robust, sustainable, and culturally attuned AI across sectors and jurisdictions.

Existing Responsible AI Maturity Models (RAIMM)

Microsoft's RAIMM, the Responsible AI Institute model, Accenture's framework, and GSMA's roadmap emphasise organizational maturity but neglect product- and project-level lifecycle alignment (Accenture, 2024; GSMA, 2024; Responsible Artificial Intelligence Institute, 2024; Vorvoreanu et al., 2023). By siloing ethical drivers, these models overlook interdependencies, which can lead to a lapse in security, undermining fairness or robustness (see Tables 3–4).

RAISEF closes these gaps by embedding privacy-preserving tools, inclusiveness, and other tailored safeguards directly into AI systems. At the same time, its cross-driver integration dynamically reconciles tensions (such as security versus fairness) and sustains both operational and ethical integrity across sectors and cultures.

Addressing all 15 drivers through three interconnected pillars, RAISEF prevents "all or nothing" failures and adjusts inter-pillar tensions to fit domain and regional contexts. The comparison in Table 4 shows how RAISEF's lifecycle integration and interactive driver typology deliver a versatile, globally applicable roadmap for Responsible AI.

Table 3

Key features and unique contributions of Responsible AI frameworks

Framework	Key Features	Limitations	RAISEF's Unique Contributions
Microsoft's RAIMM	Maturity model for Responsible AI with a focus on governance and accountability mechanisms.	Primarily organizational-level; limited sectoral or lifecycle-specific guidance.	Introduces lifecycle alignment, addressing driver interactions dynamically across development, deployment, and monitoring stages.
Google's AI Principles	Ethical principles focused on safety, fairness, privacy, and societal benefit.	High-level and abstract; lacks actionable guidance for implementation.	Provides actionable toolkits for operationalizing principles, tailored to diverse industries and cultural contexts.
EU's Ethics Guidelines	Emphasizes transparency, fairness, and trustworthiness in AI.	Primarily regulatory; lacks flexibility for regional or sectoral variations.	Combines regulatory alignment with adaptable governance frameworks, enabling implementation across sectors and jurisdictions.
IBM's AI Ethics Framework	Focus on technical robustness, transparency, and bias mitigation.	Limited focus on cross-pillar interactions or emergent properties.	Theoretical emphasis on emergent properties (e.g., trustworthiness) resulting from synergistic interactions between inclusiveness, accountability and robustness.
GSMA AI Ethics Guidelines	AI ethics tailored for the telecom sector, focusing on privacy and transparency.	Sector-specific and non-transferable; lacks lifecycle orientation.	Sector-agnostic framework adaptable to telecom, healthcare, financial services, and beyond, with explicit attention to lifecycle adaptability and scalability.

Note. Table created by the author

Table 4

Comparative analysis of RAISEF and existing Responsible AI maturity models

Feature/ Aspect	Microsoft RAIMM	Responsible AI Institute's Model	Accenture's Responsible AI Framework	GSMA's Responsible AI Maturity Roadmap	Proposed RAISEF
Focus Area	Organizational-level practices across foundations, team approaches, and Responsible AI practices	Broad organizational assessment with an emphasis on compliance and impact	Organizational and operational maturity	Industry-wide adoption in telecom	Comprehensive lifecycle integration of 15 drivers, applicable at both product/project and organizational levels
Structure	24 dimensions across three categories, with five maturity levels	General framework assessing Responsible AI readiness	Four-stage maturity framework	Structured roadmap for phased adoption	Multi-tiered model tailored to specific products, projects, and organizational goals
Industry Scope	Industry-agnostic	Industry-agnostic	Business-centric, enterprise-level focus	Telecom and mobile technology	Cross-sectoral with specific guidance for healthcare, finance, education, and more
Unique Features	Focus on empirical practices to improve organizational maturity	Strong emphasis on compliance tools and impact measurement	Focuses on aligning business operations and strategy	Designed for telecom with guidance from the AI for Impact Taskforce	Tailored toolkits for lifecycle stages, cross-cultural adaptability, and product/project-level flexibility
Inclusion of Cultural Factors	Limited	Limited	Not explicitly addressed	Not explicitly addressed	Explicit focus on cultural, regional, and sectoral adaptations to ensure global relevance

Feature/ Aspect	Microsoft RAIMM	Responsible AI Institute's Model	Accenture's Responsible AI Framework	GSMA's Responsible AI Maturity Roadmap	Proposed RAISEF
Lifecycle Integration	Limited; focuses more on organizational maturity than on AI lifecycle phases	General assessment without detailed lifecycle focus	Targets organizational integration rather than lifecycle alignment	Roadmap provides sequential adoption guidance but no lifecycle specificity	Full integration across AI lifecycle (design, development, testing, deployment, and monitoring) at product/project levels
Practical Guidance	Provides general recommendations for organizations	Offers compliance and impact assessment tools	Provides operational and organizational strategies	Offers high-level guidance specific to telecom	Detailed, step-by-step toolkits and checklists tailored to lifecycle stages and product/project-specific challenges
Originality	Builds on best practices and empirical research	General maturity assessment for Responsible AI	Collaborative development with Stanford University	Roadmap structure for adoption	Combines 15 drivers into a novel, actionable framework that balances organizational and product-level flexibility
Validation Methodology	Empirical validation of maturity dimensions	Focused on assessing compliance impact	Collaborative validation with academic institutions	Developed collaboratively but lacks empirical validation for lifecycle stages	Proposes theoretical use cases, practical pilot studies, and tailored validations at the product/project level

Note. Table created by the author

Key Differentiators of RAISEF

The following critical features differentiate RAISEF from other Responsible AI frameworks:

- Project-Level Flexibility: RAISEF enables teams to prioritize drivers at the product or project level while advancing organization-wide Responsible AI goals.
- Lifecycle Integration: Ethical safeguards, operational integrity, and societal empowerment are embedded throughout the design and post-deployment monitoring phases, ensuring continuous alignment.
- Cross-Cultural Adaptability: The framework tailors guidance to regional needs, for instance, federated learning for GDPR compliance in the EU and inclusiveness to improve healthcare access in developing economies.
- Actionable Toolkits: Sector-specific guides enable organizations at any stage of maturity to implement Responsible AI practices effectively.
- Comprehensive Novelty: By uniting 15 drivers in a single maturity model, RAISEF bridges organizational objectives with project-specific requirements.
- Scenario-Based Validation**:** Case studies in credit scoring, healthcare, education, and other fields demonstrate RAISEF's ability to strike a balance between fairness, transparency, and robustness while mitigating ethical risks.

Together, these features position RAISEF as a globally adaptable, lifecycle-aligned framework capable of addressing diverse sectoral and cultural challenges.

Driver Relationships Topology

The fifteen Responsible AI drivers operate as an interlinked system, not isolated levers. Some reinforce one another—transparency bolsters human oversight—while others compete, notably privacy against explainability. Sound design, therefore, begins with charting these links. This section (i) defines each connection, (ii) documents empirical or theoretical evidence, (iii) classifies interaction patterns, and (iv) visualizes them. Table 5 samples the 105 pairings; Appendix 5 presents the complete listing.

RAISEF manages this complexity through three levers: prioritization (emphasizing fairness in design and safety in deployment), lifecycle alignment (where, for example, federated learning protects privacy while maintaining inclusiveness), and a flexible toolkit for sector-specific conflicts. Early, adaptive integration, as illustrated by the heat map (Table 6), reduces later rework, with privacy emerging as the most frequently contested driver.

Table 5

Examples of Responsible AI driver relationships: explanations and real-world illustrations

Drivers	Relationship	Explanation	Example
Pillar: Ethical Safeguards			
Bias Mitigation vs. Fairness	■ Tensioned	While both aim to reduce injustices, techniques for fairness (e.g., demographic parity) can sometimes contradict bias mitigation goals (Ferrara, 2024).	Ensuring demographic parity in hiring algorithms might lead to the over-representation of certain groups, raising concerns about individual fairness (Dubber et al., 2020).
Inclusiveness vs. Privacy	■ Tensioned	Privacy-preserving techniques can limit data diversity, compromising inclusiveness (d'Aliberti et al., 2024).	Differential privacy in healthcare AI might obscure patterns relevant to minority groups (d'Aliberti et al., 2024).
Pillar: Operational Integrity			
Explainability vs. Robustness	■ Tensioned	High explainability may simplify models, potentially reducing their robustness (Rudin, 2019).	Simplified credit scoring models for explainability may perform poorly under non-standard conditions (Rudin, 2019).
Safety vs. Security	■ Tensioned	Adversarial robustness efforts enhance security but may reduce safety by increasing complexity (Braiek & Khomh, 2024).	Autonomous vehicle safety protocols might focus on preventing adversarial attacks at the expense of real-world robustness (Leslie, 2019).
Pillar: Societal Empowerment			
Human Oversight vs. Transparency	■ Reinforcing	Human oversight and transparency collectively foster accountability, enhancing ethical governance in AI systems (UNESCO, 2022).	In AI-driven medical diagnostics, both drivers ensure user trust and effective oversight (Ananny & Crawford, 2018).
Sustainability vs. Trustworthiness	■ Reinforcing	Sustainability and trustworthiness together enhance long-term Responsible AI deployment, creating societal and environmental benefits (van Wynsberghe, 2021).	Implementing energy-efficient AI models can increase trust, aligning with corporate sustainability goals (Accenture, 2024).

Drivers	Relationship	Explanation	Example
Cross-Pillar Tensions			
Fairness (Ethical Safeguards) vs. Robustness (Operational Integrity)	■ Tensioned	Fairness might necessitate modifications that decrease robustness (Tocchetti et al., 2022).	Adjustments to AI models for fairness in loan approvals might reduce performance across datasets (Braiek & Khomh, 2024).
Robustness (Operational Integrity) vs. Sustainability (Societal Empowerment)	■ Tensioned	Minimizing energy consumption could compromise robustness under variable conditions (Carayannis & Grigoroudis, 2023).	Energy-efficient machine learning models may struggle with edge-case data (Braiek & Khomh, 2024).

Note. Table created by the author

Ethical Safeguards

Fairness: AI must avert disproportionate harm by satisfying both group and individual criteria—demographic parity and "similar treatment" principles (Dwork et al., 2011; Kamiran & Calders, 2012; Binns, 2018; Mehrabi et al., 2019). Counterfactual methods advance this goal, as the COMPAS bias controversy demonstrates (Kusner et al., 2017; Ferrara, 2024).

Inclusiveness: Participatory design, intersectional analysis, and digital-divide initiatives broaden stakeholder representation and access (Crawford et al., 2019; Holstein et al., 2019; Fosch-Villaronga & Poulsen, 2022; World Health Organization, 2021). System-level DEI adherence increases societal relevance, with inclusive healthcare AI already reducing disparities (Shams et al., 2023; Zowghi & Da Rimini, 2024).

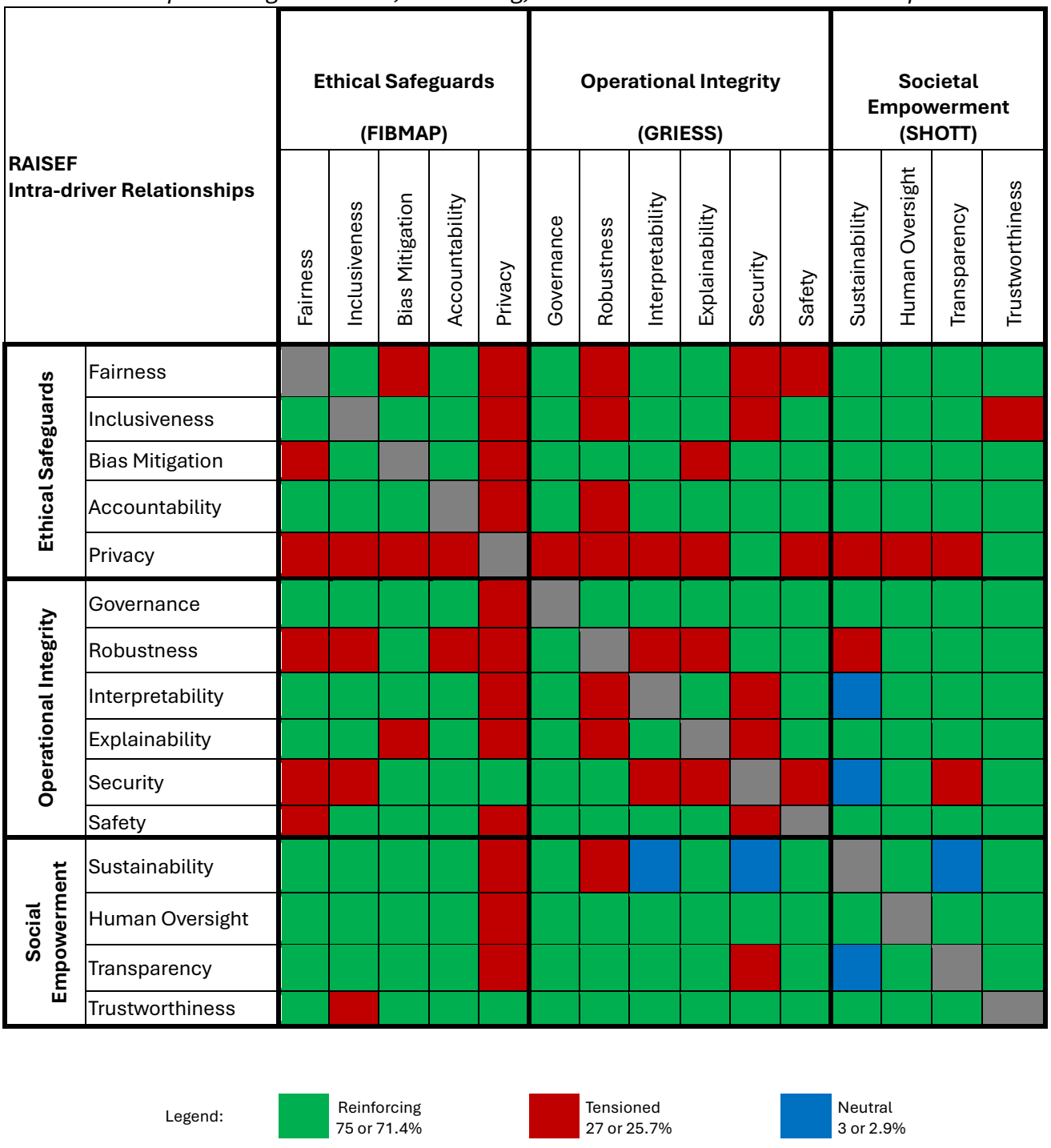
Bias mitigation: Pre-processing diversification, fairness-aware modelling, post-hoc adjustment, and adversarial defences temper prejudiced outputs, though fairness–accuracy trade-offs remain (Feldman et al., 2014; Hardt et al., 2016; Zafar et al., 2015; Zhang et al., 2018; Kleinberg et al., 2016; Danks & London, 2017). Racial misallocation in clinical risk scoring underscores the stakes (Obermeyer et al., 2019).

Accountability: Answerability, auditability, anticipatory oversight, and remedial processes anchor responsibility and public trust (Leslie, 2019; Mittelstadt, 2019; Wieringa, 2020). The Facebook–Cambridge Analytica scandal highlights the consequences of inadequate accountability structures (Novelli et al., 2024).

Privacy: Differential privacy, federated learning, homomorphic encryption, and privacy-impact assessments protect data rights and legal compliance (Dwork et al., 2011; Wachter & Mittelstadt, 2019; Kairouz et al., 2021; d'Aliberti et al., 2024). Clearview-AI litigation highlights the urgency of such safeguards (Solove, 2025).

Table 6

Heatmap showing tensioned, reinforcing, and neutral inter-driver relationships



Operational Integrity

Governance: Regulatory frameworks, ethical codes, and multi-stakeholder oversight anchor AI to societal norms, thereby closing the regulation-innovation gap (Floridi & Taddeo, 2016; Floridi et al., 2018; Fjeld et al., 2020; Gasser & Almeida, 2017). IBM's retreat from facial recognition typifies governance-driven restraint (Batool et al., 2023).

Robustness: Adversarial training, adaptive learning, and rigorous test suites sustain performance amid attacks or data drift (Goodfellow et al., 2014; Rolnick et al., 2017; Hendrycks & Dietterich, 2018; Braiek & Khomh, 2024). MRI diagnostic variance reveals the cost of weak robustness (Tocchetti et al., 2022).

Interpretability: Local and global methods, including SHAP, clarify model logic, improving clinical and policy uptake (Doshi-Velez & Kim, 2017; Ribeiro et al., 2016; Lundberg & Lee, 2017; Caruana et al., 2015).

Explainability: Post-hoc tools such as LIME justify outputs but can oversimplify; context-specific approaches temper opacity in credit scoring and other domains (Barredo-Arrieta et al., 2019; Ribeiro et al., 2016; Doshi-Velez & Kim, 2017; Rudin, 2019).

Security: Adversarial-aware models, data poisoning defenses, encryption, and automated risk management harden systems against breaches (Carlini et al., 2019; Brundage et al., 2018; Lu et al., 2024; Habbal et al., 2024). A recent healthcare intrusion underscores this need (Habbal et al., 2024).

Safety: Verification, risk controls, and fail-safes avert harm and complement robustness (Amodei et al., 2016; Goodfellow et al., 2014; Hendrycks et al., 2023). The 2018 Uber AV fatality highlights gaps in safety governance (Raji & Dobbe, 2023).

Societal Empowerment

Sustainability: Responsible AI minimises environmental harm, optimises resources, and supports equitable growth across the entire lifecycle (Rolnick et al., 2023; van Wynsberghe, 2021). Google's energy-saving data-centre algorithms exemplify this ethos, while SCAIS assesses social, ecological, and economic impacts (Rohde et al., 2023). Although critics cite higher costs and slower innovation, policy tools help reconcile efficiency with ecological goals (van Wynsberghe, 2021).

Human oversight: HITL and HOTL models keep AI aligned with human values, allowing intervention to avert harm (Rahwan, 2017; Russell, 2019; High-Level Expert Group on AI, 2019). The Dutch childcare-benefits debacle—where an algorithm falsely flagged fraud—illustrates the necessity (Bullock et al., 2024). Despite efficiency losses (Rahwan, 2017), oversight strengthens accountability and rights protection (Brundage et al., 2018).

Transparency: Clear disclosure of algorithms, interactions, and outcomes underpins accountability (Ananny & Crawford, 2018; Weller, 2017). Amsterdam's algorithm registry demonstrates practical benefits (Buijsman, 2024); however, excessive transparency can overwhelm users and compromise proprietary interests (Pasquale, 2015).

Trustworthiness: By combining fairness, accountability, and transparency, AI earns public confidence and ethical legitimacy (Toreini et al., 2019). In healthcare, trustworthy models enable accurate and explainable diagnoses; broader ethical frameworks ensure that innovation aligns with societal values, despite concerns about slowed progress (Dubber et al., 2020).

Together, these fifteen drivers enable RAISEF to deliver holistic, ethically grounded, and socially beneficial AI.

Defining Inter- and Intra-Pillar Tensions

RAISEF maps and manages frictions both within and across its three pillars. Inside Ethical Safeguards, fairness can clash with bias mitigation when data recoding shifts group equity (Ferrara, 2024). Across pillars, privacy may curb explainability because revealing model logic threatens personal data (Solove, 2025).

Healthcare diagnostics, for example, employ fairness-tailored models that benefit minority patients but compromise robustness and inclusiveness. In contrast, credit scoring relies on anonymized data that protects privacy yet hinders transparency and explanation. Table 7 visualises these tensions and synergies.

Table 7

Matrix of tensions and synergies among Responsible AI pillar drivers

Pillar/Driver	Ethical Safeguards	Operational Integrity	Societal Empowerment
Ethical Safeguards	■ Fairness vs. Bias Mitigation	■ Privacy vs. Explainability	■ Accountability vs. Trustworthiness
Operational Integrity	■ Explainability vs. Inclusiveness	■ Robustness vs. Explainability	■ Interpretability vs. Transparency
Societal Empowerment	■ Transparency vs. Privacy	■ Security vs. Sustainability	■ Human Oversight vs. Transparency

Note. Table created by the author

By weaving every driver through the AI lifecycle, RAISEF balances regulatory, ethical, and operational goals, offering a coherent template for Responsible AI.

Critical Discussion

Life-Cycle Perspective

Responsible AI drivers must integrate across the entire lifecycle—ideation, design, development, testing, deployment, and monitoring—to keep systems ethical, robust, and socially aligned (refer to Table 2).

- Ideation/Proof of Concept: Fairness, inclusiveness, and bias mitigation are key, ensuring diverse data and stakeholder voices, and setting sustainability goals (Crawford et al., 2019; Mehrabi et al., 2019). Security begins here and remains continuous.
- Design: System architecture embeds fairness, inclusiveness, transparency, and sustainability to forestall discrimination and guide ethical, eco-responsible choices (Mehrabi et al., 2019; van Wynsberghe, 2021).
- Development: Robustness testing under varied conditions, explainability tools, and governance protocols safeguard reliability, privacy, and safety (Braiek & Khomh, 2024; Doshi-Velez & Kim, 2017; Wachter & Mittelstadt, 2019).
- Testing: Accountability frameworks, interpretability checks, stress tests, and refined bias-mitigation strategies verify ethical and technical performance (Wieringa, 2020; Ribeiro et al., 2016; Amodei et al., 2016).
- Deployment: Ongoing governance tracks compliance; transparent user communication and inclusiveness ensure broad accessibility; sustainability audits assess societal and ecological impact (Cath, 2018; van Wynsberghe, 2021).
- Monitoring: DMAIC cycles-—Define, Measure, Analyze, Improve, Control—-anchor continuous audits, human oversight, and interpretability reviews to maintain integrity over time (Monday, 2022; Russell, 2019; Siau & Wang, 2018).

Drivers can be re-prioritized by phase, product, jurisdiction, or stakeholder need. Highly regulated finance may highlight transparency during deployment, while public-sector AI may enhance inclusiveness in monitoring.

In healthcare AI, for example, oversight measures outcomes, analyzes gaps, improves parameters, and controls guardrails, producing equitable, reliable, and sustainable diagnostics (Measure-Analyze-Improve-Control sequence) for present and future use.

RAISEF as a Cross-Disciplinary Framework

RAISEF blends governance, sociology, systems thinking, and design to tackle Responsible AI dilemmas. Its pillars map lifecycle interdependencies, reconciling the tensions between, for example, privacy and explainability, as well as fairness and robustness. The framework aligns with transparency, accountability, and laws such as the EU AI Act and Canada’s AIDA (Office of the European Union, 2024; Government of Canada, 2023).

Embedding equity in modular pipelines, RAISEF supports high-risk sectors such as healthcare and autonomous systems while maintaining public trust. Its theory-practice fusion equips scholars, policymakers, and practitioners to address AI's ethical, operational, and societal demands.

Interrelatedness of Drivers

Responsible AI drivers form an integrated system. RAISEF charts 15 drivers and 105 pairwise links, highlighting their mutual influence. Fairness depends on bias mitigation because bias distorts outcomes across groups (Mehrabi et al., 2019). Inclusiveness deepens bias work by adding diverse perspectives that strengthen fairness and increase transparency (Crawford et al., 2019; Holstein et al., 2019). Such multivocal scrutiny clarifies decision logic, bolstering explainability, interpretability, and trust (Doshi-Velez & Kim, 2017; Ribeiro et al., 2016). Accountability ties these elements together by holding actors answerable for system outputs and aligning practice with RAISEF's pillars (Wieringa, 2020). Overlaps thus create synergies, not redundancies.

Governance translates accountability into enforceable norms, supporting fairness and safety (Brundage et al., 2018). Robustness and safety co-reinforce reliability across conditions, sustaining public confidence (Amodei et al., 2016), while security protects both robustness and privacy from unauthorized access (Brundage et al., 2018). Human oversight ensures autonomous behaviour remains aligned with human values, balancing autonomy and accountability (Russell, 2019).

Sustainability links governance and inclusiveness, demanding environmental stewardship alongside equity (Rolnick et al., 2023; van Wynsberghe, 2021). Trustworthiness emerges as the cumulative outcome of these balanced drivers; without their coordinated implementation, it cannot be achieved (Toreini et al., 2019).

These interdependencies demonstrate that Responsible AI is not a checklist, but an integrated design philosophy applied throughout the system's lifecycle.

Practical Implications for Stakeholders

RAISEF's success hinges on multi-stakeholder cooperation. Leaders must embed the 15 drivers into policy and culture, striking a balance between profit and societal impact (Dubber et al., 2020; Rahwan, 2017). Regulatory alignment follows from collaboration with policymakers and targeted staff training in ethical AI (Russell, 2019). Policymakers should craft adaptive, accountable governance (Batool et al., 2023); data scientists must pair innovation with bias mitigation and robustness (Mehrabi et al., 2019); end-users should demand transparency for informed trust (Doshi-Velez & Kim, 2017). Robust accountability, inclusive design, multidisciplinary teams, privacy-preserving methods, human oversight, and continuous monitoring operationalize these aims (Crawford et al., 2019; Holstein et al., 2019; Rahwan, 2017). Scaling horizontally across project phases and vertically through organizational tiers, RAISEF offers a practical roadmap for ethical, innovative, market-aligned AI.

Case Studies and Their Role in Demonstrating RAISEF

The following case studies are hypothetical, scenario-based illustrations of how RAISEF can be applied across sectors; they are not empirical evaluations.

Case Study 1: Healthcare diagnostics. This scenario applies RAISEF to a retinal-imaging triage tool used across urban hospitals and rural clinics. The case: speed up referrals while preserving clinician trust. The challenge: under-representation of certain patients, privacy constraints on image sharing, and the tension between explainability and performance. The resolution: build a representative training set (plus carefully governed synthetic augmentation), add differential privacy on sensitive fields, and require human-in-the-loop review for high-risk calls. Clinician-facing explanations (saliency overlays, short reason codes) and model cards support accountability. Continuous monitoring tracks subgroup sensitivity/specificity, calibration, and escalation rates, with a fairness threshold agreed by a clinical governance board. A lightweight RAISEF scorecard helps weigh accuracy, equity, and privacy impacts over time.

Case Study 2: Credit scoring in the EU. A mid-size lender uses RAISEF to expand access while meeting GDPR and consumer-protection rules. The case: responsibly incorporate alternative data to reach thin-file applicants. The challenge: potential proxy bias, strict data-minimization, and the right to explanation/appeal. The resolution: limit features to necessity, apply privacy-preserving processing, use fairness-aware training with monotonic constraints, and deliver clear adverse-action reason codes. Human reviewers handle borderline decisions; applicants get a documented appeal path. Post-deployment, an independent auditor validates drift, reject-inference procedures, and disaggregated default rates. A public dashboard reports stability, calibration, and equal-opportunity metrics; an internal RAISEF dashboard quantifies trade-offs among approval lift, risk, and fairness.

Case Study 3: Smart agriculture in resource-constrained settings. A smallholder advisory platform in Sub-Saharan Africa uses RAISEF to guide planting and irrigation recommendations. The case: deliver equitable advice despite sparse labels, variable connectivity, and diverse microclimates. The challenge: data scarcity and potential harm from one-size-fits-all guidance. The resolution: combine satellite and weather data with locally crowdsourced observations; run lightweight models on low-cost phones; provide SMS/IVR in local languages; and embed extension-worker review for high-impact decisions. Fairness checks compare recommendation quality across soil types and farm sizes; seasonal cross-validation guards against climate shift. Monitoring tracks water-use efficiency, yield proxies, and uptake, with red-team tests for unintended impacts. A RAISEF scorecard helps prioritize equity, sustainability, and reliability.

Case Study 4: Smart cities and urban governance. A mid-sized North American city coordinates AI for traffic optimization, benefits pre-screening, and place-based risk analysis under RAISEF. The case: improve services while preserving rights. The challenge:

balancing privacy and transparency, avoiding disparate impact, and governing cross-agency data sharing. The resolution: co-design with community groups; use privacy-preserving aggregation and retention limits; require human review for adverse eligibility decisions; and publish plain-language model cards. Quarterly audits report disaggregated error rates by neighborhood and demographic groups; appeal portals and an ethics board enforce accountability. Interoperable logs support independent oversight. A municipal RAISEF dashboard tracks service quality, fairness, and complaint resolution, guiding continuous improvement.

Empirical pilots and a RAISEF Scorecard (weighted metrics, driver benchmarks, sector thresholds) are planned to validate these scenarios and provide quantitative decision-support.

Diagnostics, credit scoring, smart agriculture, interdisciplinary culture, and urban governance cases demonstrate RAISEF's sector-agnostic versatility. The framework embeds ethical safeguards and operational integrity throughout the AI lifecycle. In healthcare (Case Study 1), it reconciles HIPAA privacy with explainability through interdisciplinary teams, securing equitable outcomes for underserved groups.

European credit scoring (Case Study 2) strikes a balance between fairness and accountability, leveraging inclusiveness to mitigate bias while ensuring transparency and compliance for equitable access to credit. Whereas, Sub-Saharan smart agriculture (Case Study 3) merges sustainability and inclusiveness, reconciling needs with global ecological goals through adaptive solutions.

Smart cities (Case Study 4) evidence RAISEF's value in multi-stakeholder public services. Applied to municipal-AI, such as predictive policing, welfare screening, and traffic optimization, it balances fairness, privacy, and transparency in high-impact contexts. Participatory design, open data, and lifecycle audits embed governance and human oversight, sustaining public trust and community rights while advancing data-driven innovation.

These cases demonstrate RAISEF's 15 drivers reconcile privacy-explainability and security-safety conflicts, embedding Responsible AI across the lifecycle and fortifying ethical resilience.

Challenges and Limitations

Persistent gaps in measurement, regulation, and stakeholder alignment hinder the development of Responsible AI. Unrepresentative data undermines fairness and efficacy, while rapid innovation outpaces law, creating an oversight "pacing problem" (Birkstedt et al., 2023). Sparse real-world data limits model generalizability (Wirtz et al., 2024), and socioeconomic disparities impede equitable deployment, especially in low-resource settings (Birkstedt et al., 2023). Conflicting interests and definitions of fairness further complicate consensus, demanding multidisciplinary collaboration and adaptive

governance (Ferrara, 2024). Frameworks that link ethical principles to practice are therefore vital for achieving equitable and effective AI across diverse sociotechnical contexts (Batool et al., 2023).

Limitations and Future Research

Despite its conceptual breadth and practical relevance, RAISEF's is tempered by key limits. Its lifecycle model lacks empirical trials, leaving effectiveness, scalability, and cultural fit unproven. Responsible AI work also faces challenges such as weak regulation, stakeholder misalignment, and sparse representative data, particularly in underserved regions (Ferrara, 2024; Wirtz et al., 2024; Dubber et al., 2020). Finally, standardized benchmarks and metrics for constructs such as trustworthiness remain undeveloped (Toreini et al., 2019; Mittelstadt, 2019).

Future research will empirically test RAISEF through a structured doctoral project. Pilot studies in AI-driven healthcare diagnostics and credit scoring in domains marked by concerns over fairness, accountability, privacy, and explainability will assess RAISEF's practical relevance and adaptability. A forthcoming RAISEF Scorecard will quantify lifecycle performance via weighted metrics, driver benchmarks, and sector thresholds. Radar and heatmap dashboards offer real-time insights into the readiness of Responsible AI across various maturity stages.

Beyond empirical trials and scorecard design, the dissertation will test RAISEF in multiple settings that intertwine AI with democratic accountability, data justice, and public trust. Empirical, technical, and policy refinements will transform RAISEF into a globally adaptable standard, resilient to risks from generative and autonomous systems, as well as evolving regulations. Future work will extend the framework to novel domains. In generative AI, RAISEF will aim to curb hallucinations, IP breaches, misinformation, and opacity, supported by auditability, curated datasets, and user disclosures. In high-risk autonomous arenas, including vehicles, industrial robotics, and defense, it will embed human oversight, safety protocols, escalation paths, and adversarial testing to uphold reliability and public trust.

These expansions build on Case Study 4, demonstrating how lifecycle-aligned guidance fosters equitable, transparent, and accountable AI in civic contexts that have a direct impact on the community.

Conclusion

Summary of Key Findings

This study introduces RAISEF, which integrates 15 drivers across three pillars: ethical safeguards, operational integrity, and societal empowerment to close gaps in Responsible AI. It resolves tensions such as privacy-explainability and fairness-robustness through actionable inter- and intra-pillar methods. Embedding oversight from ideation

through post-deployment, while accommodating diverse cultures, RAISEF averts early conflicts, enhances transparency, and secures stakeholder trust and system robustness.

RAISEF flexibility adapts to diverse cultural and regulatory settings, aligning inclusiveness with societal empowerment. It tailors regional solutions, balancing GDPR privacy requirements with the data needs of developing nations, and scales across various sectors. Healthcare applications reconcile confidently with interpretability; financial services secure fair, robust credit scoring; and agricultural deployments deliver transparent AI that benefits marginalized smallholder farmers.

Futureproofing RAISEF

As AI advances, responsible frameworks must evolve. RAISE remains pertinent by mapping ethical oversight across the entire system lifecycle and organizational layers. Generative models such as GPT-4o spur creativity yet raise concerns about accountability, fairness, and misinformation (Bommasani et al., 2021). RAISEF embeds fairness, transparency, robustness, and bias mitigation to curb misuse.

Autonomous systems, ranging from self-driving cars to critical-care AI, require frameworks balancing safety, oversight, and inclusion. Tesla's FSD disputes highlight why robust safety protocols and accountability matter. RAISEF's human-centric governance pillar scales oversight to autonomy level and risk (Raji & Dobbe, 2023). Amid the EU AI Act, Canada's AIDA, and China's directives (European Parliament, 2017; Government of Canada, 2023), RAISEF reconciles regional nuances with universal standards, offering sector-specific guidance for global compliance.

By integrating emerging technological trends and global regulations, RAISEF remains a dynamic framework that evolves in tandem with AI, providing resilient theoretical and practical guidance for future Responsible AI challenges.

Contributions to Knowledge and Practice

RAISEF embeds ethical principles across the AI lifecycle, connecting organizational strategy to product-level practice, drawing on ethics, governance, and systems thinking. This study introduces tools, such as inter-driver conflict matrices and lifecycle-prioritization mechanisms, to navigate Responsible AI trade-offs. RAISEF accommodates diverse cultural and regulatory contexts, empowering stakeholders to build ethical, sustainable, and globally scalable AI.

Encouraging Adoption and Pilots of RAISEF

Stakeholders should pilot RAISEF to validate and refine it. A forthcoming practitioner book will guide the adoption of Responsible AI. RAISEF.ai will host current sector-specific resources, such as examples, templates, checklists, case studies, and implementation guides, to help organizations of all technical maturity levels deploy RAISEF. Collaborative tools enable stakeholders to share insights, while periodic updates from pilot projects iteratively refine the framework for emerging industry demands.

Visit https://raisef.ai for the case studies and appendices mentioned in this paper.

References

Accenture. (2024). From compliance to confidence: Embracing a new mindset to advance responsible AI maturity. https://www.accenture.com/us-en/insights/data-ai/compliance-confidence-responsible-ai-maturity

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety [Preprint]. arXiv. http://arxiv.org/abs/1606.06565

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media and Society, 20(3), 973–989. https://doi.org/10.1177/1461444816676645

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI [Preprint]. arXiv. https://arxiv.org/abs/1910.10045

Batool, A., Zowghi, D., & Bano, M. (2023). Responsible AI governance: A systematic literature review [Preprint]. arXiv. http://arxiv.org/abs/2401.10896

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In S. A. Friedler & C. Wilson (Eds.), In Proceedings of the 1st conference on Fairness, Accountability, and Transparency (pp. 149–159). PMLR. http://proceedings.mlr.press/v81/binns18a/binns18a.pdf

Birkstedt, T., Minkkinen, M., Tandon, A., & Mäntymäki, M. (2023). AI governance: themes, knowledge gaps and future agendas. Internet Research (Vol. 33, Issue 7, pp. 133–167). https://doi.org/10.1108/INTR-01-2022-0042

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). On the opportunities and risks of foundation models [Preprint]. arXiv. https://arxiv.org/abs/2108.07258

Braiek, H. Ben, & Khomh, F. (2024). Machine learning robustness: A primer [Preprint]. arXiv. http://arxiv.org/abs/2404.00897

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., Ó Éigeartaigh, S., Beard, S., Belfield, H., Farquhar, S., … Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation [Preprint]. arXiv. https://arxiv.org/abs/1802.07228

Buijsman, S. (2024). Transparency for AI systems: a value-based approach. Ethics and Information Technology, 26(2). https://doi.org/10.1007/s10676-024-09770-w

Bullock, J. B., Chen, Y.-C., Himmelreich, J., Hudson, V. M., Korinek, A., Young, M. M., & Zhang, B. (Eds.). (2024). The Oxford handbook of AI governance. Oxford University Press. https://academic.oup.com/edited-volume/41989

Carayannis, E. G., & Grigoroudis, E. (Eds.). (2023). Handbook of research on artificial intelligence, innovation and entrepreneurship. Edward Elgar Publishing Limited.

Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., & Kurakin, A. (2019). On evaluating adversarial robustness [Preprint]. arXiv. http://arxiv.org/abs/1902.06705

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015-August, 1721–1730. https://doi.org/10.1145/2783258.2788613

Cath, C. (2018). Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. In Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences (Vol. 376, Issue 2133). Royal Society Publishing. https://doi.org/10.1098/rsta.2018.0080

Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., Kak, A., Mathur, V., McElroy, E., Sánchez, A. N., Raji, D., Rankin, J. L., Richardson, R., Schultz, J., West, S. M., & Whittaker, M. (2019). AI Now 2019 report. AI Now Institute. https://ainowinstitute.org/wp-content/uploads/2023/04/AI_Now_2019_Report.pdf

d'Aliberti, L., Gronberg, E., & Kovba, J. (2024). Privacy-enhancing technologies for artificial intelligence-enabled systems [Preprint]. arXiv. http://arxiv.org/abs/2404.03509

Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. In Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017). https://www.cmu.edu/dietrich/philosophy/docs/london/IJCAI17-AlgorithmicBias-Distrib.pdf

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning [Preprint]. arXiv. http://arxiv.org/abs/1702.08608

Dubber, M. D., Pasquale, F., & Das, S. (Eds.). (2020). The Oxford handbook of ethics of AI. Oxford University Press. https://academic.oup.com/edited-volume/34287

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). Fairness through awareness [Preprint]. arXiv. http://arxiv.org/abs/1104.3913

European Parliament. (2017). The AI act (Directive TA-8-2017-0051). http://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.pdf

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2014). Certifying and removing disparate impact [Preprint]. arXiv. http://arxiv.org/abs/1412.3756

Ferrara, E. (2024). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. Sci, 6(1), 3. https://www.mdpi.com/2413-4155/6/1/3

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. C., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI (Berkman Klein Center Research Publication No. 2020-1). https://doi.org/10.2139/ssrn.3518482

Floridi, L., & Taddeo, M. (2016). What is data ethics? In Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences (Vol. 374, Issue 2083). Royal Society of London. https://doi.org/10.1098/rsta.2016.0360

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Fosch-Villaronga, E., & Poulsen, A. (2022). Diversity and inclusion in artificial intelligence. In B. Custers & E. Fosch-Villaronga (Eds.), Law and Artificial Intelligence (pp. 109–134). https://doi.org/10.1007/978-94-6265-523-2_6

Gasser, U., & Almeida, V. A. F. (2017). A layered model for AI governance. IEEE Internet Computing, 21(6), 58–62. https://doi.org/10.1109/MIC.2017.4180835

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples [Preprint]. arXiv. https://arxiv.org/abs/1412.6572

Google. (2023). Google AI principles. https://ai.google/responsibility/principles/

Government of Canada. (2023). The artificial intelligence and data act (AIDA): Companion document. https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document

GSMA. (2024). The GSMA responsible AI maturity roadmap. https://www.gsma.com/solutions-and-impact/connectivity-for-good/external-affairs/wp-content/uploads/2024/09/GSMA-ai4i_The-GSMA-Responsible-AI-Maturity-Roadmap_v8.pdf

Habbal, A., Ali, M. K., & Abuzaraida, M. A. (2024). Artificial intelligence trust, risk and security management (AI TRiSM): Frameworks, applications, challenges and future research directions. Expert Systems with Applications (Vol. 240). https://doi.org/10.1016/j.eswa.2023.122442

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems, 3323–3331. https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

Hendrycks, D., & Dietterich, T. G. (2018). Benchmarking neural network robustness to common corruptions and surface variations [Preprint]. arXiv. http://arxiv.org/abs/1807.01697

Hendrycks, D., Mazeika, M., & Woodside, T. (2023). An overview of catastrophic AI risks. arXiv. http://arxiv.org/abs/2306.12001

High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

High-Level Expert Group on Artificial Intelligence. (2020). Assessment list for trustworthy artificial intelligence (ALTAI) for self-assessment. European Commission. https://doi.org/10.2759/791819

Holstein, K., Vaughan, J. W., Daumé, H., Dudík, M., & Wallach, H. (2019, May 2). Improving fairness in machine learning systems: What do industry practitioners need? In Proceedings of the Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3290605.3300830

IBM. (2024). IBM AI ethics. https://www.ibm.com/impact/ai-ethics

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., … Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning (Vol. 14, Issues 1–2, pp. 1–210). Now Publishers Inc. https://doi.org/10.1561/2200000083

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores [Preprint]. arXiv. http://arxiv.org/abs/1609.05807

Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). Counterfactual fairness [Preprint]. arXiv. http://arxiv.org/abs/1703.06856

Leslie, D. (2019). Understanding artificial intelligence ethics and safety. https://doi.org/10.5281/zenodo.3240529

Leslie, D., Rincón, C., Briggs, M., Perini, A., Jayadeva, S., Borda, A., Bennett, S., Burr, C., Aitken, M., Katell, M., Fischer, C., Wong, J., & Garcia, I. K. (2024). AI fairness in practice. The Alan Turing Institute. https://www.turing.ac.uk/sites/default/files/2023-12/aieg-ati-fairness_1.pdf

Lu, Q., Zhu, L., Whittle, J., & Xu, X. (2024). Responsible AI: Best practices for creating trustworthy AI systems. Pearson Education. https://www.pearson.com/en-us/subject-catalog/p/responsible-ai-best-practices-for-creating-trustworthy-ai-systems/P200000010211/9780138073886

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions [Preprint]. arXiv. https://arxiv.org/abs/1705.07874

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning [Preprint]. arXiv. http://arxiv.org/abs/1908.09635

Microsoft. (2022). Microsoft responsible AI standard, v2. https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

Monday, L. M. (2022). Define, measure, analyze, improve, control (DMAIC) Methodology as a roadmap in quality improvement. Global Journal on Quality and Safety in Healthcare, 5(2), 44–46. https://doi.org/10.36401/jqsh-22-x2

Novelli, C., Taddeo, M., & Floridi, L. (2024). Accountability in artificial intelligence: What it is and how it works. AI and Society, 39(4), 1871–1882. https://doi.org/10.1007/s00146-023-01635-y

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://www.science.org/doi/10.1126/science.aax2342

Office of the European Union. (2024). Artificial Intelligence Act, The. http://data.europa.eu/eli/reg/2024/1689/oj

Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Harvard University Press. https://www.hup.harvard.edu/books/9780674970847

Rahwan, I. (2017). Society-in-the-loop: Programming the algorithmic social contract. Ethics and Information Technology, 20(2018), 5–14. https://doi.org/10.1007/s10676-017-9430-8

Raji, I. D., & Dobbe, R. (2023). Concrete problems in AI safety, revisited [Preprint]. arXiv. https://arxiv.org/abs/2401.10899

Responsible Artificial Intelligence Institute. (2024). Our responsible AI maturity model. https://www.responsible.ai/our-responsible-ai-maturity-model

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-August-2016, 1135–1144. https://doi.org/10.1145/2939672.2939778

Rohde, F., Wagner, J., Meyer, A., Reinhard, P., Voss, M., Petschow, U., & Mollen, A. (2023). Broadening the perspective for sustainable AI: Sustainability criteria and indicators for artificial intelligence systems [Preprint]. arXiv. https://arxiv.org/abs/2306.13686

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A. S., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C. P., Ng, A. Y., Hassabis, D., Platt, J. C., … Bengio, Y. (2023). Tackling climate change with machine learning. ACM Computing Surveys, 55(2), 1–96. https://doi.org/10.1145/3485128

Rolnick, D., Veit, A., Belongie, S., & Shavit, N. (2017). Deep learning is robust to massive label noise [Preprint]. arXiv. http://arxiv.org/abs/1705.10694

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence (Vol. 1, Issue 5, pp. 206–215). Nature Research. https://doi.org/10.1038/s42256-019-0048-x

Russell, S. (2019). Human compatible artificial intelligence and the problem of control. https://doi.org/10.1007/978-3-030-86144-5_3

Russell, S., & Norvig, P. (2021). Artificial intelligence: A modern approach (4th ed.). Pearson. https://www.pearson.com/en-ca/subject-catalog/p/artificial-intelligence-a-modern-approach/P200000003500/9780134610993

Shams, R. A., Zowghi, D., & Bano, M. (2023). AI and the quest for diversity and inclusion: A systematic literature review. AI and Ethics. https://doi.org/10.1007/s43681-023-00362-w

Siau, K. L., & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. Cutter Business Technology Journal, 31(2), 47–53. https://www.researchgate.net/publication/324006061

Solove, D. J. (2025). Artificial intelligence and privacy. Florida Law Review, 77(1), 1–73. https://doi.org/10.2139/ssrn.4713111

Tocchetti, A., Corti, L., Balayn, A., Yurrita, M., Lippmann, P., Brambilla, M., & Yang, J. (2022). A.I. robustness: A human-centered perspective on technological challenges and opportunities [Preprint]. arXiv. https://arxiv.org/abs/2210.08906

Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C. G., & van Moorsel, A. (2019). The relationship between trust in AI and trustworthy machine learning technologies [Preprint]. arXiv. http://arxiv.org/abs/1912.00782

UNESCO. (2022). Recommendation on the ethics of artificial intelligence. https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence

van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. AI and Ethics, 1(3), 213–218. https://doi.org/10.1007/s43681-021-00043-6

Vorvoreanu, M., Heger, A., Passi, S., Dhanorkar, S., Kahn, Z., & Wang, R. (2023). Responsible AI maturity model: Mapping your organization's goals on the path to responsible AI (Microsoft White Paper). Microsoft. https://www.microsoft.com/en-us/research/uploads/prod/2023/05/RAI_Maturity_Model_Aether_Microsoft_whitepaper.pdf

Wachter, S., & Mittelstadt, B. (2019). A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. Columbia Business Law Review (Vol. 2019). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3248829

Weller, A. (2017). Transparency: Motivations and challenges [Preprint]. arXiv. http://arxiv.org/abs/1708.01870

Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT), 1–18. https://doi.org/10.1145/3351095.3372833

Wirtz, B. W., Langer, P. F., & Weyerer, J. C. (2024). An ecosystem framework of AI governance. In J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson, A. Korinek, M. M. Young, & B. Zhang (Eds.), The Oxford handbook of AI governance. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197579329.013.24

World Health Organization. (2021). Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. https://www.who.int/publications/i/item/9789240029200

Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2015). Fairness constraints: Mechanisms for fair classification [Preprint]. arXiv. http://arxiv.org/abs/1507.05259

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning [Preprint]. arXiv. http://arxiv.org/abs/1801.07593

Zowghi, D., & Da Rimini, F. (2024). Diversity and inclusion in artificial intelligence. In Q. Lu, L. Zhu, J. Whittle, & X. Xu (Eds.), Responsible AI: Best practices for creating trustworthy AI systems (Chap. 11). Pearson Education. https://www.pearson.com/en-us/subject-catalog/p/responsible-ai-best-practices-for-creating-trustworthy-ai-systems/P200000010211/9780138073886