The Building Blocks of a Responsible Artificial Intelligence Practice: An Outlook on the Current Landscape

Maryem Marzouk, *Euranova R&D, Tunis, 2015, Tunisia*

Cyrine Zitoun, Oumaima Belghith, and Sabri Skhiri, *Euranova R&D, Mont-Saint-Guibert, 1435, Belgium*

For artificial intelligence (AI)-driven companies, awareness of the urgency of the responsible application of AI became essential with increased interest from different stakeholders. Responsible AI (RAI) has emerged as a practice to guide the design, development, deployment, and use of AI systems to ensure a benefit to users and those impacted by the systems' outcomes. This benefit is achieved through trustworthy models and strategies that assimilate ethical principles to ensure compliance with regulations and standards for long-term trust. However, RAI comes with the challenge of lack of standardization when it comes to which principles to adopt, what they mean, and how they can be operationalized. This article aims to bridge the gap between principles and practice through a study of different approaches taken in the literature and the proposition of a foundational framework.

The rapid evolution of artificial intelligence (AI) has unlocked an era of technological advancement impacting diverse sectors from health care and finance to transportation and entertainment. Its potential to transform traditional processes, enhance efficiency, and promote new possibilities is driving worldwide and cross-industry adoption. However, alongside the promise lie complex ethical and regulatory concerns.[1] Questions related to transparency, bias, privacy, and control have emerged at the forefront of the AI discourse. The convergence of these ethical dilemmas with legal constraints creates a landscape where the responsible development, use, and governance of AI are imperative. Addressing these challenges is crucial to exploiting AI's transformative potential.

Responsible AI (RAI) has emerged as a key governance framework that ensures adherence to ethical principles and best practices, thereby fostering a trustworthy environment for AI deployment, scaling, and innovation.

This article demystifies the concept of trust in AI, examining practical, actionable methods for the operationalization of ethics. We propose well-articulated criteria to answer the following research questions (RQs):

- "What": What characterizes an RAI framework?
- "How": How can RAI frameworks successfully translate ethical considerations into practical operations?

The "Research Methodology" section starts with a review of prominent RAI frameworks. We present it through an elaborate assessment of common objectives, features, and tools that support principles' integration. This analysis triggered the need to review predominant principles in the extensive literature. To this end, we propose selection criteria to tackle ambiguity and the issues of lacking standards hindering the adoption of principles. Our assessment approach filters a substantial number of RAI principles to arrive at a compendious selection addressing the different dimensions of RAI. This selection drives trust in AI while ensuring transparency, fairness, robustness, and accountability and establishing the link between data and AI governance.

The overarching ambition of this study is to augment the existing body of knowledge tackling RAI in practice, infusing it with a unique analytical perspective.

1541-1672 © 2023 IEEE
Digital Object Identifier 10.1109/MIS.2023.3320438
Date of publication 29 September 2023; date of current version 13 November 2023.

The findings derived from our research methodology were used to build a comprehensive metamodel. This foundational framework offers a simplified, yet profound understanding of the prevailing disconnect between theoretical principles and tangible practices. It contributes with different approaches to closing that gap, backed by extensive research, and justified by our own criteria.

RESEARCH METHODOLOGY

Despite it being a relatively nascent practice (see "Appendix A" in the supplementary downloadable material available at https://10.1109/MIS.2023.3320438), there is an abundance of RAI resources, guidelines, and tools developed and made available by companies, practitioners, researchers, and regulators. Yet, the transition from principles to practice remains a challenge.

In its earlier years, RAI emerged with the proposal of principles[2] to drive trust in AI. The principles, however, proved to be too ambiguous, conceptual, and numerous to be adopted and operationalized. These issues have been the center of attention of many researchers in AI ethics and RAI.

Notable works include "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI,"[3] where the authors assess guidelines for AI that have emerged from various initiatives. The report analyzes trends to establish a shared understanding of ethical principles, revealing eight key themes: privacy, accountability, safety and security, transparency and explainability, fairness, human control of technology, professional responsibility, and the promotion of human values.

A more comprehensive review is provided in "The Global Landscape of AI Ethics Guidelines,"[4] which focuses on resolving issues of diverging interpretations, varying significance, issue pertinence, application domain, and implementation means for RAI principles. It concludes with a thematic analysis of five ethical principles: transparency, justice and fairness, nonmaleficence, responsibility, and privacy. The authors highlight the need for more harmonization and standardization of such guidelines to ensure their effectiveness and global adoptability.

In another significant study, "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices,"[2] the authors highlight the need for more research to evaluate the effectiveness of AI ethics tools and propose an *applied ethical AI typology* to outline where research efforts should be focused and the role that AI practitioners need to play in the advancement of ethical AI.

To address the remaining unanswered questions, we started with a thorough study of practical initiatives presented through an elaborate analysis of 14 RAI frameworks. For principles, we performed a three-phase sampling process on more than 300 principles, which resulted in 15 principles that are actionable, measurable (quantifiable), and verifiable (integrated in RAI assessments).

Our criteria are exhaustively presented in the next sections.

This systematic review was conducted by accessing a variety of search engines using general search keywords combining "AI" or "artificial intelligence" with "responsible," "ethical," "trustworthy," and next with "principles," "guidelines," "ethics," "frameworks," "applications," "toolkit," and "tools." We also utilized open access repositories and web portals[a,b,c] covering RAI publications and guidelines.

As a final step, we surveyed related works and existing literature reviews and inspected the websites of private companies offering developed frameworks.

All the publications found through referencing and citation chaining were checked to make sure their content complied with our criteria.

This review provides inclusive coverage in terms of the nature of issuing parties according to which frameworks and principles were selected. These referential natures are industry-related issuers, governmental issuers, agencies, and intergovernmental organizations, and, finally, academia, nonprofits, and nongovernmental organizations.

A Review of RAI Frameworks

The systematic review, described previously, resulted in 26 initial frameworks focused on RAI. This result was narrowed down, and the final selection was conducted according to specific criteria, with a set of assessment criteria for the analysis.

Selection Criteria

The following selection criteria were used:

- *Practice pertinence*: Frameworks designed for AI governance without incorporating ethical considerations and technical requirements are out of scope (OOS).
- *Frameworks based on a set of principles*: Those built around only one principle are OOS.

[a] LAIP: Linking Artificial Intelligence Principles, by the Institute of Automation, Chinese Academy of Sciences.
[b] AIethicist.org.
[c] AI Ethics Guidelines Global Inventory, by Algorithm Watch.

- *Frameworks applicable to all AI systems, projects, and practice domains*: Those intended for a particular AI application or use case are OOS.
- *Frameworks for civilian applications of AI*: Military applications are OOS.
- *Frameworks applicable to narrow AI*[d] *systems*: Those governing artificial general intelligence and artificial superintelligence are OOS.

This first combination of selection criteria yielded 21 RAI models. However, for a more decisive analysis, we applied the following requirements:

1) Framework practicality.
   a) *Governmental use*: Frameworks intended to guide governments, policymakers, regulators, standard-setting and oversight bodies, or organizations developing AI for the exclusive use of these entities.
   b) *Business use*: Frameworks intended to
      i) Guide businesses by providing open access models for unguided implementation or self-assessment.
      ii) Or explain a company's internal approach to RAI.
      iii) Or provide a guided consulting solution.
   c) *Cross* practical.
2) *Actionable recommendations, strategy proposals, and leading practices*: (i.e., detailed and applicable in real business or policy settings).
3) *Tools*: Framework support through technical toolkits, nontechnical methodologies and conceptual frameworks, or assessments.

These requirements produced 14 RAI frameworks. We propose the full overview resulting from the complete sampling process in "Appendix B" in the supplementary downloadable material available at https://10.1109/MIS.2023.3320438, with an emphasis on the ones compliant with our criteria and requirements.

Assessment Approach

Although each one is unique in its approach, we identified common properties among the 14 selected frameworks. The following sections define these properties and present excerpts of their analysis through examples.

Assessment According to Objectives

The objectives of a framework answer the following questions: How does the framework approach RAI? and how do its tools and recommendations integrate principles? The applicability to respective principles needs to be explicit.

Principles-Specific Objectives

The following principles-specific objectives were identified:

1) *Operationalizing principles*: (e.g., technical/practical checks, algorithmic tools, governance strategies, or decision-making assistance).
2) *Quantifying principles*: (e.g., the methods used to quantify trust through the implementation of the principle, its impact on the outcomes of the system, or the level of risk associated with it).
3) *Qualifying principles*: (e.g., clear definitions, or the qualitative evaluation of risk, impact, or opportunity associated with principles).

General Objectives Based on Approach

The following general objectives were identified:

4) *RAI assessment*: Which is used to determine the degree of alignment of AI governance processes or system functionalities, design, and performance with ethical principles and responsibility guidelines.
5) *RAI governance*: The organization of roles, responsibilities, and processes to structure and monitor AI throughout its lifecycle.
6) *Provision of practice recommendations.*

Table 1 includes examples with different practicalities for insight into structural differences deduced from objectives. The frameworks that are practical for governmental use share RAI assessments as their main focus, while their integration of principles converges toward qualification or quantification. However, those designed for business use focus on operationalizing principles and balance between RAI assessment, RAI governance, and the provision of practice recommendations. See "Appendix C" in the supplementary downloadable material available at https://10.1109/MIS.2023.3320438 for the full assessment.

Assessment According to Main Features

Main features are summarized in the following assessment properties:

1) *Application scope*: Generalizable/ use case specific/customizable.
2) *Key focus*: Organizational/operational/technical/societal/systems oriented.
3) *RAI enablers*.
4) *Built to integrate RAI throughout all AI lifecycle processes.*

[d] *Narrow AI* refers to AI systems designed to perform specific and limited tasks.

TABLE 1. Excerpt of the assessment according to objectives.

	Objectives					
Framework	**1**	**2**	**3**	**4**	**5**	**6**
PricewaterhouseCoopers (business practicality)	☑	—	—	☑	☑	☑
AI Ethics Impact Group (governmental practicality)	—	☑	—	☑	—	—
The European Commission's High-Level Expert Group on Artificial Intelligence (cross practical)	—	—	☑	☑	☑	☑

The examples in Table 2 represent different application scopes and key foci to illustrate the different features. See "Appendix D" in the supplementary downloadable material available at https://10.1109/MIS.2023.3320438 for a full assessment.

The main takeaway is the identification of possible RAI enablers in an organizational setting. These enablers cover four main axes: people, processes, technology, and organizational culture.

Assessment According to Tools

This step in the assessment was conducted across the following properties:

1) *Purpose*: Operational or assessment tools.
2) *Accessibility*: Proprietary/available for public access.
3) *The means to operationalize principles.*
4) *Principles addressed.*
5) *RAI assessment tools.*

   a) *Self-/guided assessment.*
   b) *The means to quantify principles.*
   c) *Rating scale and metrics.*
   d) *Actionable recommendations following the assessment.*

The "Appendix E" in the supplementary downloadable material available at https://10.1109/MIS.2023.3320438 contains a full analysis against these properties. To summarize, we distinguished between eight frameworks whose tools are exclusively focused on RAI assessment, four that offer tools for RAI operationalization, and two that fulfill both. This differentiation between technical tools and practical methods that support the development, assessment, and use of AI reveals an overlap between business and technology approaches. This offers a unique perspective into how effective RAI can be as supported by the examples of IBM,[5] Microsoft,[6] and Google.[7]

Although we cannot describe all the analyzed technical tools in detail, we introduce how they support frameworks and discuss current limitations.

In this context, we present a discussion of the main tools in Microsoft's RAI toolkit, which provide a large, functional coverage and represent state-of-the-art tools. Nonetheless, they still have significant limitations.

Fairlearn

This toolkit assesses and mitigates whether AI models perform as well for one person or another. It evaluates

TABLE 2. Excerpt of the assessment according to main features.

Framework	**Application scope**	**Key focus**	**RAI enablers**	**All lifecycle processes**
AI Ethics Impact Group	Generalizable	Systems oriented	N/A	Yes
IBM	Use-case specific	Organizational + technical	RAI governance Continuous education and guidance	Yes
ITechLaw	Customizable	Societal + organizational	Governance People Process Technology	Yes

N/A: not applicable.

the fairness principle. The toolkit is composed of 1) a fairness assessment and 2) fairness mitigation tools.

With the fairness assessment, developers can analyze false-positive and false-negative rates with regard to sensitive attributes but can also observe the disparity in terms of prediction. The fairness mitigation tool leverages exponentiated-gradient reduction[8] for generating a sequence of relabeling, reweighting, and training a predictor for each. The optimization process takes a fairness metric, such as the demographic parity that constrains predictions to be independent of sensitive attributes, to generate different model versions. These versions offer different tradeoffs between accuracy and fairness (discrepancy of predictions).

At the time of writing this article, Fairlearn is available for tabular data and for classification and regression models. However, text, graph, and image data are not yet supported.

InterpretML

This toolkit exposes interpretation algorithms through a unified application programming interface (API). The package provides two classes of models: 1) a black-box model API exposing algorithms such as LIME[9] or SHAP[10] and 2) a glass-box model API exposing algorithms that are explainable by nature, such as decision trees, linear models, and the explainable boosting model.[11] Although it provides a nice abstraction model on top of explainability models, it still has limitations. For example, in Kumar et al.,[12] the authors highlight the mathematical limitation of different approximation techniques such as SHAP or LIME, human-centric issues, and the difficulty with causal reasoning. Lipton[13] describes the limitation of local interpretability. Finally, Zhang et al.[14] argue that in structure-aware problems like graphs, SHAP is suboptimal.

Error Analysis

This toolkit allows 1) identification of cohorts with a high error rate versus benchmarking and visualization of how the error rate is distributed and 2) diagnosing the root causes of the errors by visually diving deeper into the characteristics of data and models (via its embedded interpretability capabilities). The first use enables the model designer to understand the regions of the data distribution where the model is less successful, while the second integrates explainable methods of InterpretML for understanding why the model is underperforming. As a result, as diagnostic tools reuse InterpretML, the same limitations bind them. Notice that error analysis also includes Dice[15] for a counterfactual explanation. However, it is still limited to classifiers, mainly on tabular data.

Counterfit

This toolkit was one of the first on the market for assessing AI risks and estimating vulnerabilities. The current version is mainly focused on adversarial attacks, but it also provides support for text, tabular data, and images. Counterfit leverages several open source projects such as ATLAS,[16] Adversarial Robustness Toolbox,[5] and TextAttack.[17] Although this is a remarkable tool in the AI security landscape, it still has limited coverage. Within Microsoft’s published threat taxonomy listing the possible attacks and failures of AI models, Counterfit only addresses adversarial attacks.

In conclusion, RAI tools are still in a nascent stage. They are an excellent first attempt to equip data scientists with powerful tools to ensure trustworthiness, explainability, and fairness. However, their scope is still highly limited.

A Review of RAI Principles

Almost all existing publications and frameworks regulating, discussing, or implementing RAI are based on principles. Consequently, an examination of these principles is needed to complement our discussion of practical frameworks.

In this review, we propose criteria that allow for the positioning of the current landscape of RAI principles. We then conclude with principles that are most suited for integration in a framework.

This review does not aim to cover the comprehensive landscape of RAI principles. Instead, we limit our focus to the results of our criteria. More overarching reviews have been introduced by other studies.[3,4]

Selection Criteria and Approach

This selection offers a contribution to existing research through the identification of four primary focus areas in most discussions of principles: putting principles to practice, policymaking, standard setting, and ethical focus.

The selection by primary focus areas, coupled with the diverse natures of issuing entities, highlighted principles from 45 documents spanning the past six years. This step revealed that there is no consensus on the designation of principles. One principle can have up to nine different designations in different documents. Additionally, a principle having the same designation can have varying definitions and interpretations depending on the source and its primary focus area. The next section aims to resolve these issues.

Following a filtering process, we selected principles according to common designations and gradually classified them according to treated issues relevant to technology, social impact, and ethics.

To arrive at a more decisive assessment, we applied the following requirements:

- *The definition of each principle in its source*: Principles are classified according to the definition/interpretation toward which most publications converge.
- *Focus area*: Putting principles to practice is the area of interest.
- *Sampling principles from frameworks previously reviewed*: This is applicable only when the principle is explicitly operationalized.

In total, nine out of 45 documents discussing principles met the selection criteria and requirements (see "Appendix F" in the supplementary downloadable material available at https://10.1109/MIS.2023.3320438). Along with the classified principles from frameworks, we identified a total of 33 general principles.

Assessment Approach

The objective of this assessment is to categorize all 33 general principles under qualifiable, quantifiable, and operationalizable core principles, described through subprinciples and defined through indicators.

This final categorization was conducted according to

1) *The definition of each principle in its source and the links it reveals to other principles.*
2) *Similarities in target focus and treated issues.*
3) *The ability of the core principle to capture the theme of the group.*
4) *Trends in the reviewed literature.*

   a) *Trends in the citation count*: Trends are spotted in the principles of fairness, transparency, accountability, privacy, explainability, security, reliability, and data governance.
   b) *Trends in grouping and categorizing principles*: These trends are illustrated in a conceptual map in "Appendix G" of the supplementary downloadable material available at https://10.1109/MIS.2023.3320438. Designing this graph consisted of an analysis of common groupings of multiple principles per category and aimed to reveal otherwise-concealed links through the number of occurrences of the same group in multiple sources.

This tracking was key to the final categorization of four core principles, each described through a set of subprinciples, summarized as follows (see "Appendix H" in the supplementary downloadable material available at https://10.1109/MIS.2023.3320438 for complete definitions):

1) *Transparency*: AI transparency is achieved through reliable, auditable, and traceable documentation of the information about AI systems and related business processes.
   a) *Explainability and interpretability*: This principle focuses on technical transparency in terms of the explanation of model mechanisms and predictability of system outcomes.
   b) *Auditability*: Defined as the readiness of the organization and its AI systems for a review of algorithms, data, design, and processes.
2) *Data governance*: This core principle allows for control over the data feeding AI models in terms of quality, reliability, and accessibility.
   a) *User data rights*: Which focus on the provision of mechanisms allowing data subjects control over their personal data and awareness of their usage and processing.
   b) *Privacy*: Which is made possible through the application of privacy by default and privacy by design[e] as well as protection against data breaches and privacy threats.
3) *Responsibility*: Which focuses on the provision of trusted AI-powered products and services and the establishment of an organizational RAI culture.
   a) *Fairness*: Which is resumed in the avoidance of creating or reinforcing unfair bias inherent in the data as well as ensuring diversity in datasets and AI teams to ensure equal treatment.
   b) *Accountability*: Which clarifies who in the organization is responsible for AI failure and who is liable in cases of negative consequences. Accountability is ensured through prospective, monitoring, as well as contestability and redress measures.
   c) *Autonomy*: Which is resumed in the exercise of human agency and oversight following clearly defined model governance mechanisms.
   d) *Lawfulness and compliance*: Which covers regulatory and legal compliance and accordance with relevant industry standards and best practices.

[e] Subject of Article 25 of the General Data Protection Regulation. Privacy should be integrated into the products, processes, and systems, from the early stages of the design through their lifecycle and organizations (data controller) must ensure that only data strictly necessary for each specific purpose are processed without the intervention of the user (data subject).

4) *Robustness*: Which focuses on the prevention and minimization of unintentional harm, safeguarding against adversarial threats and attacks, meeting performance requirements, and operating as intended.

    a) *Safety and security*: Which focuses on the prevention and minimization of harm resulting from internal malfunctioning of AI systems (safety) and safeguarding against external threats (security).
    b) *Reliability*: Which focuses on meeting performance requirements, accounting for uncertainty, and operating as intended. Outputs of AI systems must be trustworthy (i.e., accurate, reproducible, able to handle exceptions, and fit for purpose).
    c) *Monitoring and moderation*: Which calls for continuous testing, validation, and monitoring to ensure robustness and fallback planning in cases of failure or harm.

RAI Metamodel

To frame the full findings of our research and analysis, we developed an all-encompassing metamodel that aims to establish the building blocks for the development of an RAI practice.

The classification of surveyed documents under the metamodel is based on each referenced initiative's affinity with the objectives identified in the "Research Methodology" section. We performed a comprehensive mapping of each document to its applicable principles and each principle to target objectives.

The graphical visualization is enclosed in "Appendix I" of the supplementary downloadable material available at https://10.1109/MIS.2023.3320438 and provides an illustrative view of our research, steered by its criteria and guided by its outputs.

Most importantly, the metamodel highlights the convergence of all features of selected frameworks, research papers, governmental resources, and academic contributions toward our set of selected principles.

These principles crown the metamodel edge to show that ethics initiate trust in AI. Looking inward, it shows the method through which the selected initiatives treat these principles and how they approach RAI in practice.

The objectives are ordered from specific to general, starting with the qualification, quantification, and operationalization of principles, leading to RAI assessment, governance, and contribution through practice recommendations.

This methodology addresses the gaps between principles and practice that surround RAI by showcasing how they are addressed in the literature.

Although simple in its structure, the metamodel is meant to provide readers, at a glance, with insights into the initiatives it covers, and consequently, into the RAI landscape as a whole and the challenges of complexity and variability still surrounding it.

Primarily, it shows the RAI principles around which attention is centered. Although these trends were utilized in our selection, mapping principles to objectives reinforces their predominance. We can observe concentrated attention on fairness, transparency, accountability, privacy, data governance, safety and security, and reliability. Consequently, the metamodel tackles 1) the complexity challenge by providing an overview of how our simplified selection indeed covers the most important dimensions of RAI and 2) the variability challenge by demonstrating the consensus that converges toward those dimensions (in varying degrees for each principle and objective).

Particularly, general objectives mapped to principles can answer the question, "How are these principles being operationalized?" If a document is classified under operationalization and practice recommendations or governance, that means that those recommendations or practices partially or wholly answer this question. This is done to uncover overlapping objectives.

To elaborate on the structural differences highlighted by the metamodel, it extends an outlook into the general structure followed by each initiative and whether it follows that same structure for all of its principles.

To illustrate, the AI Ethics Impact Group[18] uniformly addresses the principles of reliability, transparency, fairness, accountability, and privacy through qualification and quantification. In contrast, PricewaterhouseCoopers's[1] approach differs. All the principles are addressed through qualification and quantification, except for the principle adaptation and moderation, which is only qualified.

Additionally, we identified a general convergence toward fairness, explainability and interpretability, and accountability when it comes to the objective of operationalization. This objective is of high relevance to this article and to the future work it aims to support.

Other relevant centers of attention from an objectives frame of reference can be spotted in fairness, accountability, and explainability and interpretability through providing practice recommendations, while governance mechanisms are centered around accountability. Accountability is the focal point of governance because many entities assign ethics committees as an accountability measure. Their responsibilities, however, extend beyond accountability to enable and oversee all other principles and address ethical concerns.

Another purpose of this metamodel is to offer a tool that inspires and supports practitioners undertaking an RAI initiative and developing RAI frameworks. It helps in understanding the current state of ethics and trust in practice to build on the highlighted approaches of mapped principles to objectives.

DISCUSSION

The starting point of our research was to introduce and analyze RAI frameworks originating from industry, academia, governments, and governmental organizations.

The criteria and approach followed to assess the findings led to structured conclusions and empirical insight. This insight should grant readers an understanding of imperatives that make an effective RAI framework and the different approaches to conducting structured assessments, applying principles, and strategizing AI governance.

Next, we focused on RAI principles as the common integral constituent of all RAI discussions: from ethics, to technology, to practice. For this, we sampled principles while focusing on

- sorting through variabilities in definitions and interpretations. (e.g., differentiating between explainability and interpretability).
- untangling complexities. (e.g., using the conceptual map in "Appendix G" of the supplementary downloadable material available at https://10.1109/MIS.2023.3320438 as a tool to provide readers with a better-informed understanding of the literature).
- contributing with a proposition of principles to adopt.

This sampling process and the resulting selection do not disregard principles that are not explicit in the final categorization, but rather define core principles through them. Hence, classification through addressed issues proved to be the appropriate methodology to ensure that all the major dimensions of RAI are treated in our principles, and in turn, in the metamodel.

Structuring the metamodel also relied heavily on the frameworks' analysis and their assessment against objectives. It offers insights into RAI as a practice and could serve as a tool to benchmark future initiatives using the same parameters and taxonomy.

Admittedly, from a principles standpoint, the model depicts a lack of applicability of several principles to certain objectives. For example, it shows no reference to methodologies for the quantification of user data rights, auditability, and adaptation and moderation. In addition, autonomy, lawfulness and compliance, and auditability are overlooked when it comes to operationalization. However, we conducted the frameworks analysis according to generalizability, which aims to close these gaps by identifying applicable generalizable methodologies to quantify, assess, and operationalize principles following a flexible, broad-ranging structure.

With the formalization of our definitions of the metamodel's RAI principles, an important conclusion to draw is how to define the concept of trust in AI. Our findings confirm that trust can be achieved only through the complementation of organizational and technical capabilities. From a technical perspective, trust in AI is illustrated in our definition of reliability, which is the notion of correctness and trust in the outputs of machine learning models. However, robustness through reliability goes hand in hand with transparency through the explainability of those outputs. Therefore, trust is enabled by both the technical robustness and explainability of the models[19] and complemented by responsible and accountable business measures.

CONCLUSION

This survey presents excerpts of the findings of a larger research effort to address RAI practices. Starting with the criteria presented throughout the article, we focused on RAI frameworks that contribute with actionable practices and comprehensive features and tools. Thus, answering RQ1: "What characterizes an RAI framework?"

We then studied discussions of principles that share putting principles into practice as the key focus area.

As an end goal, we introduced the metamodel as a tool to position the current landscape and suggested means to adopt and scale AI safely, transparently, and responsibly. This foundational framework answers RQ2: "How can RAI frameworks successfully translate ethical considerations into practical operations?"

We urge RAI practitioners to utilize the approaches to RAI highlighted by the metamodel in conjunction (e.g., RAI assessments are useful only if we use their results to enforce effective governance and implement tools and practices. The same reasoning applies to all combinations of objectives).

Although there are other relevant works that were not reviewed in this survey, the RAI frameworks and principles compiled and analyzed are prime examples of the current state of RAI in practice. They therefore assist in framing the context of Euranova's tailor-made RAI framework. This framework is consistent with state-of-the-art practices, aligned with the insights derived from our metamodel, fixated on bridging gaps revealed by this research, and built on in-house expertise in the fields of data governance, consulting, and AI engineering. It is intended to assist organizations with an end-to-end assessment of their systems and governance, define road maps to guide RAI journeys through their evaluation, the definition of strategies, and validation of RAI programs.

ACKNOWLEDGMENT

This work is part of a larger R&D effort initiated by Euranova to introduce its own RAI framework of reference. Through its research projects, Euranova aims to develop cutting-edge services providing a competitive advantage to its clients. This article has supplementary downloadable material available at https://10.1109/MIS.2023.3320438.

REFERENCES

1. “PWC’s responsible AI toolkit,” PricewaterhouseCoopers. [Online]. Available: https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai.html
2. J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, “From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices,” *Sci. Eng. Ethics*, vol. 26, no. 4, pp. 2141–2168, Dec. 2019, doi: 10.1007/s11948-019-00165-5.
3. J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, “Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI,” *SSRN Electron. J.*, early access, Jan. 2020, doi: 10.2139/ssrn.3518482.
4. A. Jobin, M. Ienca, and E. Vayena, “The global landscape of AI ethics guidelines,” *Nature Mach. Intell.*, vol. 1, no. 9, pp. 389–399, Sep. 2019, doi: 10.1038/s42256-019-0088-2.
5. B. Green, D. Heider, K. Firth-Butterfield, and D. Lim, *Responsible Use of Technology: The IBM Case Study*. Cologny/Geneva, Switzerland: World Economic Forum, 2021.
6. “Responsible AI resources.” Microsoft AI. [Online]. Available: https://www.microsoft.com/en-us/ai/responsible-ai-resources
7. “Building responsible AI for everyone.” Google AI. Accessed: Sep. 1, 2022. [Online]. Available: https://ai.google/responsibilities/
8. A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, “A reductions approach to fair classification,” in *Proc. 35th Int. Conf. Mach. Learn.*, PMLR, 2018, vol. 80, pp. 60–69.
9. M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you?” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
10. H. Nori, S. Jenkins, P. Koch, and R. Caruana, “InterpretML: A unified framework for machine learning interpretability,” 2019, *arXiv:1909.09223*.
11. Y. Lou, R. Caruana, J. Gehrke, and G. Hooker, “Accurate intelligible models with pairwise interactions,” in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 623–631, doi: 10.1145/2487575.2487579.
12. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, “Problems with Shapley-value-based explanations as feature importance measures,” in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 5491–5500.
13. Z. C. Lipton, “The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery,” *Queue*, vol. 16, no. 3, pp. 31–57, May/Jun. 2018, doi: 10.1145/3236386.3241340.
14. S. Zhang, Y. Liu, N. Shah, and Y. Sun, “GStarX: Explaining graph neural networks with structure-aware cooperative games,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2022.
15. R. K. Mothilal, A. Sharma, and C. Tan, “Explaining machine learning classifiers through diverse counterfactual explanations,” in *Proc. Conf. Fairness, Accountability, Transparency*, 2020, pp. 607–617, doi: 10.1145/3351095.3372850.
16. “Mitre/Advmlthreatmatrix: Adversarial threat landscape for AI systems.” GitHub. Accessed: Sep. 1, 2022. [Online]. Available: https://github.com/mitre/advmlthreatmatrix
17. “QData/TextAttack: TextAttack is a python framework for adversarial attacks, data augmentation, and model training in NLP.” GitHub. Accessed: Sep. 1, 2022. [Online]. Available: https://github.com/QData/TextAttack
18. “From principles to practice: An interdisciplinary framework to operationalize AI ethics,” AI Ethics Impact Group, Apr. 2020. [Online]. Available: https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WKIO_2020_final.pdf
19. A. Holzinger, “The next frontier: AI we can really trust,” in *Proc. ECML PKDD, Commun. Comput. Inf. Sci.*, M. Kamp, Ed., Cham, Switzerland: Springer Nature, 2021, pp. 427–440, doi: 10.1007/978-3-030-93736-2_33.

MARYEM MARZOUK is a data governance officer at Euranova R&D, Tunis, 2015, Tunisia. Her research interests include artificial intelligence (AI) ethics, AI strategy and data management. Marzouk received her BSBA degree in business analytics and IT from Tunis Business School. Contact her at maryem.marzouk@euranova.eu.

CYRINE ZITOUN is a data governance consultant at Euranova R&D, Mont-Saint-Guibert, 1435, Belgium. Her research interests include data management, governance, and artificial intelligence ethics. Zitoun received her master's degree in business analytics from South Mediterranean University-Mediterranean school of business. Contact her at cyrine.zitoun@euranova.eu.

OUMAIMA BELGHITH is a data governance consultant at Euranova R&D, Mont-Saint-Guibert, 1435, Belgium. Her research interests include data and information management and governance. Belghith received her BSBA in business analytics and finance from Tunis Business School. Contact her at oumaima.belghith@euranova.eu.

SABRI SKHIRI is the R&D director at Euranova, Mont-Saint-Guibert, 1435, Belgium, and the chief visionary officer for Digazu, Brussels, Belgium. His research interests include high scalability, cloud and elasticity, and data management. Skhiri received his master's degree in computer science engineering from the University of Brussels. Contact him at sabri.skhiri@euranova.eu.