Translating Responsible AI Principles into Practice: Insights from a Pilot Project

Eleonora Viganò | Eugenia Cacciatori | Christian Hauser

Summary

Current Responsible AI (RAI) frameworks provide high-level ethical principles, but the development of practical procedures and tools for effective business implementation remains limited. Our pilot project "From ethical principles to practical implementation: Exploring the challenges of consulting in the implementation of RAI" investigated how Swiss organizations approach RAI implementation, explored their key challenges and opportunities, and established a foundation for developing practical solutions in a subsequent project aimed at narrowing the gap between RAI theory and business practice.

Eleven semi-structured interviews and a workshop revealed that:

- Organizations view RAI as synonymous with ethical AI but interpret "ethical" through different lenses
- Organizations lack a clear point of responsibility for RAI initiatives
- The main obstacles to RAI implementation are:
    - Limited awareness of its significant costs
    - Difficulty in defining and translating RAI components into metrics and practices
    - A "wait-and-see" attitude
    - Limited top management endorsement
    - Regulatory ambiguity
- The primary drivers of RAI implementation are:
    - Awareness of AI risks
    - Top management support
    - Regulatory pressure
- Organizations seek modular solutions adaptable to different use cases, industries, and daily workflows, including tools for measuring implementation progress
- Organizations anticipate significant growth in RAI services, tools, and skills

Based on these findings, we aim to develop a RAI Toolkit in collaboration with industry partners. This toolkit will be tailored to specific organizational use cases and AI technologies for integration into business operations. Our approach ensures the toolkit becomes an actionable tool for RAI implementation, embedded in organizational governance and daily decision-making processes. This represents an initial step toward developing modular solutions.

1. Introduction

The rapid advancement of AI technologies has heightened concerns about their ethical implications. AI systems present unique ethical challenges compared to traditional software: they may interact with individuals without explicit knowledge, often remain opaque even to their designers, and might incorporate significant biases [1], [2]. In response, governments and international organizations are developing regulatory frameworks to mitigate potential negative impacts. The EU AI Act will significantly influence many Swiss companies' AI practices. Meanwhile, various organizations, consulting companies, and scholars are developing frameworks beyond mere compliance, focusing on Responsible AI (RAI) principles. While definitions vary, RAI fundamentally means developing, deploying, and using AI in ways that uphold ethical values such as fairness, transparency, and accountability [3]. Most existing RAI frameworks provide high-level ethical principles but lack practical procedures and tools for effective business implementation [1], [2], [4], [5], [6]. This creates challenges for companies trying to integrate ethical considerations into their workflow. Moreover, RAI implementation must be adapted to specific contexts, industries, and company departments, adding further complexity. Our pilot project "From ethical principles to practical implementation: Exploring the challenges of consulting in the implementation of RAI" investigated how consultancy organizations and their clients approach RAI implementation, explored key challenges and opportunities, and established groundwork for developing practical solutions. This report presents the main findings based on eleven semi-structured interviews and a workshop conducted in English. These findings will inform the development of a project proposal on RAI implementation in collaboration with partner companies in the second research phase.

2. Methods

We conducted eleven interviews with RAI experts providing consultancy to companies (n=2) and employees working in RAI-related fields across consultancy companies (n=7) and organizations using AI (n=2). Interviewees represented diverse Swiss entities: small ethics consultancies, an international consulting company's Swiss branch, a large data service provider, a non-profit organization, an AI compliance service provider, and a publicly owned company. This qualitative research aimed to capture interviewees' experiences and perspectives on implementing RAI.

A subsequent workshop, "Consulting in RAI: How to Narrow the RAI Implementation Gap", was held at the University of Zurich on September 2, 2024. The workshop involved thirteen participants: nine from six Swiss companies and four from academic institutions. The workshop validated interview findings and generated ideas for practical solutions for RAI implementation. Thematic content analysis was used on the interview and workshop transcripts.

3. Results

The thematic content analysis revealed six primary themes regarding Swiss organizations' approaches to RAI implementation:

3.1 Understanding of RAI

Organizations consider responsible AI synonymous with ethical AI development and use. However, they conceptualize "ethical" differently: as robust, fair and explainable; as respecting data protection, individual autonomy, and transparency; as trustworthy; as respecting fairness, privacy, security, and transparency; as developing good, trustworthy, and reliable systems; and as creating positive societal impact. This broad consistency in understanding, coupled with internal diversity, offers both opportunities and challenges. It enables construction of shared interest but risks breaking down when priorities and inevitable trade-offs require management.

3.2 Lack of Clear Responsibility

Organizations struggle to identify specific individuals responsible for RAI initiatives and budget allocation. AI and digital technology responsibilities typically disperse across Digital Responsibility teams, IT, and Legal departments.

> *"At the moment there are different people and different roles who are working or will work on this topic [RAI]. Of course, the legal department is involved, especially to monitor all the regulations, and our IT department is also involved"*

3.3 Implementation Challenges

High costs and limited awareness: The cost of implementing RAI emerged as the primary challenge, as companies often lack dedicated RAI budgets and underestimate the costs and time involved. This includes some companies' belief they are already doing enough for RAI while they do not have even minimal AI ethics safeguards (e.g., checking for toxic chatbot responses). One interviewee expressed concern that RAI can create trade-offs with business objectives, risking override in favor of business goals. Thus, RAI implementation may incur both direct and opportunity costs.

> *"Certain checks and regulations could potentially impact the viability of certain projects. If I were a business manager with a strong desire to see a project through, I might consider ways to circumvent such checks"*

RAI definition and translation: Organizations struggle to define and translate ethical concepts and principles into metrics and actionable operations. Interviewees express uncertainty about the practical meaning of privacy, transparency, trustworthiness, and responsibility. Technical staff mentioned difficulty finding metrics for measuring AI project trustworthiness, while management staff from consultancies struggled to apply ethical pillars to specific client solutions.

> *"There's a real proliferation of principles and manifestos and cartas and documents [on RAI]. [...] But what does it mean in your daily work? [...] It's really difficult to tell people what [RAI] means and to operationalize it"*

Wait-and-see attitude: Companies' hesitant approach toward implementing RAI presents another obstacle, both as a psychological tendency and a characteristic of Swiss business culture.

> *"It's humans themselves with their hesitation wanting to wait and see what is happening. But this is normal human behavior, especially in Switzerland to just wait and see what the others are doing"*

Lack of top management support: Without top management endorsement, RAI practices risk being deprioritized or overlooked.

> *"If there's not a top-down decision and support from the management to really do this [implement RAI], then it will not be very sustainable"*

Regulatory ambiguity: Some companies view regulations, particularly the EU AI Act, as obstacles due to their vagueness and insufficient guidance. However, most companies view regulations as RAI drivers, as we will see in the next section.

> *“We know that [the EU AI Act] is still vague, particularly for certain industries. When you look at the framework of the EU AI Act, it is horizontal and geared towards customer protection and individual human rights”*

3.4 Facilitators of RAI Implementation

Several companies' facilitators mirror the hindrances discussed above. The most cited facilitator is awareness of AI risks, particularly the implications for individuals and society.

> *“Organizations need to be aware that they need to implement AI responsibly, some only see the opportunities but don't see the risks”*

Proposed awareness-building strategies include employee training, discussions on AI impact, RAI certifications, and initiatives such as the World Economic Forum. Companies view awareness as necessary but not sufficient to implement RAI, requiring complementary top management commitment to make RAI a strategic priority and ensuring influence on companies' operations for those responsible for RAI. Most consider regulatory pressure an enabler and accelerator of RAI implementation, prompting companies to view RAI as a mitigator of legal and reputational risks.

> *“Today, one of the strongest driving forces behind discussions on RAI is the upcoming regulation of this technology. This impending regulation creates a clear need for organizations in certain sectors to start preparing for compliance”*

3.5 Desired RAI Implementation Tools

The workshop confirmed that organizations primarily struggle with translating abstract RAI principles into concrete actions within specific contexts. Participants expressed strong interest in developing versatile, modular solutions adaptable to specific user needs and use cases, particularly focusing on fairness and accessibility. The most supported solutions included developing industry-specific best practices, creating Key Performance Indicators (KPIs) for RAI, and introducing risk management frameworks with clear achievement thresholds.

3.6 The Future RAI Market

Interviewees unanimously predicted significant expansion in RAI services and tools, particularly AI governance tools and RAI certifications. One interviewee characterized this anticipated growth as a transition “from virtually zero to the size of the privacy market”. However, this rapid market expansion raises potential challenges. Some warned about the risk of certification and tool proliferation and difficulties in quality assessment, noting that not all RAI consultancy providers would offer high-quality products grounded in comprehensive understanding of ethical principles, societal contexts, and technical nuances. More optimistic perspectives emerged as well, anticipating stabilization and homogenization of RAI offerings and vocabularies. Some predicted clearer understanding of responsible AI requirements, potentially aided by educational developments like machine learning (ML) courses.

> *“I think at some point the ML AI classes will start teaching the toxicity metric and then maybe it will become a more known thing how to measure it”*

4. Conclusions and Follow-ups

The interviews and workshop highlight the need for actionable and quantifiable processes to implement high-level RAI frameworks. Such processes must integrate into organizational and strategic processes to address trade-offs between ethical considerations and business goals. Based on these findings, we plan to partner with organizations to seek funding from sources such as Innosuisse to develop a scientifically grounded RAI Toolkit based on specific use cases and AI technologies.

We envision two research lines for developing this toolkit. The first addresses responsibility and transparency in multi-actor AI systems, aiming to create a prototype toolkit offering a structured approach to responsibility. This includes developing procedures for assessing social, moral, and environmental responsibilities, establishing criteria for responsibility allocation across organizational roles and departments, and formulating guidelines for responsibility mitigation. Our prototype toolkit would feature an organizational mapping component that enables companies to visualize and enhance the flow of RAI-related decisions and information within their organization. This would allow companies to address obstacles to RAI implementation such as the unclear responsibilities identified in our research.

The second research line focuses on providing companies with advanced benchmarking tools (e.g., RAI-KPIs) for each established RAI component, enabling effective measurements of ethical and compliance performances in AI development and deployment, alongside a model supporting progressive advancement in ethical governance. The RAI-KPIs would enable the development of a modular approach in which each step delivers immediate, measurable value.

The RAI Toolkit will offer a structured approach to help bridge the implementation gap. The project's findings will provide an empirically grounded procedure for translating high-level RAI principles into practical business workflows, effectively bridging theoretical concepts with operational realities.

Project Team

Project Leader
Eleonora Viganó
University of Zurich and University of Applied Sciences of the Grisons

Lead Team
Eugenia Cacciatori
Bayes Business School, City St George's, University of London

Christian Hauser
University of Applied Sciences of the Grisons

Eleonora Viganó
University of Zurich and University of Applied Sciences of the Grisons

Team Members
Jana Sedlakova
University of Zurich

Andrea Ferrario
University of Zurich

Stewart Palmer
Aarhus University

Christian Hugo Hoffmann
Swiss AI Startup Center and University of Zurich

References

[1] V. Vakkuri, K.-K. Kemell, and P. Abrahamsson, “AI ethics in industry: a research framework,” arXiv preprint arXiv:1910.12695, 2019.

[2] J. Krijger, T. Thuis, M. de Ruiter, E. Ligthart, and I. Broekman, “The AI ethics maturity model: a holistic approach to advancing ethical data science in organizations,” AI and Ethics, vol. 3, no. 2, pp. 355-367, 2023.

[3] V. Dignum, “Ensuring Responsible AI in Practice,” in Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way, V. Dignum, Ed., Cham: Springer International Publishing, 2019, pp. 93-105. doi: 10.1007/978-3-030-30371-6_6.

[4] J. Ayling and A. Chapman, “Putting AI ethics to work: are the tools fit for purpose?,” AI and Ethics, vol. 2, no. 3, pp. 405-429, Aug. 2022, doi: 10.1007/s43681-021-00084-x.

[5] J. C. Ibáñez and M. V. Olmeda, “Operationalising AI ethics: how are companies bridging the gap between practice and principles? An exploratory study,” AI & Society, vol. 37, no. 4, pp. 1663-1687, 2022.

[6] R. Eitel-Porter, “Beyond the promise: implementing ethical AI,” AI and Ethics, vol. 1, no. 1, pp. 73-80, 2021.

Acknowledgements
Funding by the Ethics Community of the Digital Society Initiative (University of Zurich), the Innovation Booster Databooster (Innosuisse), and the companies participating in the workshop is gratefully acknowledged. We wish to thank the interviewees and workshop participants for generously contributing their time to this research. The report was edited with assistance from Claude by Anthropic.

ISBN: 978-3-9524437-4-3
FH Graubünden Verlag, Chur, 2024

Contact for requests:
eleonora.vigano@uzh.ch

Cover photo: Adobe Stock, INTERPIXELS