

June 2025

## How to Operationalize Responsible Use of Artificial Intelligence

Lorenn P. Ruster

Katherine A. Daniell

Follow this and additional works at: <https://aisel.aisnet.org/misqe>

---

### Recommended Citation

Ruster, Lorenn P. and Daniell, Katherine A. (2025) "How to Operationalize Responsible Use of Artificial Intelligence," *MIS Quarterly Executive*: Vol. 24: Iss. 2, Article 6.  
Available at: <https://aisel.aisnet.org/misqe/vol24/iss2/6>

This material is brought to you by the AIS Journals at AIS Electronic Library (AISeL). It has been accepted for inclusion in MIS Quarterly Executive by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact [elibrary@aisnet.org](mailto:elibrary@aisnet.org).

# How to Operationalize Responsible Use of Artificial Intelligence

*The explosion in AI development, application and implementation has led to increasingly urgent calls for the responsible use of AI—as well as increasing confusion about how to practically approach it. This article describes how two organizations started to operationalize responsible AI by taking a systems approach to crafting responsibility pledges and embedding them in organizational practices. We identify a five-phase process, along with corresponding activities and artifacts, and share effectiveness evidence and a roadmap for action.<sup>1,2</sup>*

**Lorenn P. Ruster**

School of Cybernetics, Australian National  
University (Australia)

**Katherine A. Daniell**

School of Cybernetics, Australian National  
University (Australia)

## Introduction

Amid the recent artificial intelligence (AI) boom, there has been a corresponding call to build and deploy responsible AI. The need to comply with laws such as the EU AI Act is one part of the responsible AI puzzle and can be largely accommodated through existing organizational approaches to risk management and regulatory affairs.<sup>3</sup> How to act on other parts of the responsible AI puzzle, however, is less clear; specifically, how organizations can embrace proactive, responsible action.

Responsible AI can take a variety of forms. Indeed, there is no universal definition of responsible AI, let alone how to operationalize it successfully. For this article, we provide the following definition of responsible AI: “The practices undertaken by humans to design, develop, deploy and govern AI systems in ways that adhere to fundamental values and benefit individuals, groups, wider society and the environment, while minimizing the risk of negative consequences.”<sup>4</sup>

<sup>1</sup> Hind Benbya is the accepting senior editor for this article.

<sup>2</sup> The authors thank the senior editor and two anonymous reviewers for their constructive feedback and thoughtful guidance throughout the review process. The authors thank members of the participating organizations who generously enabled this research. The authors are also appreciative of the encouragement of Alja Isaković and Mathew Mytka from Tethix, who have been trialing responsibility pledges in their own work. The authors would also like to thank William J. Kettinger and three anonymous reviewers who gave valuable feedback on an earlier iteration of this paper submitted to the Hawaiian International Conference on Systems Sciences (HICSS). This research was supported by a Florence Violet McKenzie scholarship and an Australian Government Research Training Scholarship.

<sup>3</sup> The EU AI Act is the world’s first comprehensive law on regulating AI and, as such, is being considered as a model for various other jurisdictions, such as Australia. The Act takes a risk-based approach to regulating AI use, providing obligations depending on risk level. See *EU AI Act: First Regulation on Artificial Intelligence*, European Parliament, August 6, 2023, available at <https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>.

<sup>4</sup> Adapted from Vassilakopoulou, P., Parmiggiani, E., Shollo, A. and Grisot, M. “Responsible AI Concepts, Critical Perspectives and an Information Systems Research Agenda,” *Scandinavian Journal of Information Systems* (34:2), 2022, pp. 89-112; and Responsible AI Pattern Catalogue, Commonwealth Scientific and Industrial Research Organisation, 2024, available at <https://research.csiro.au/ss/science/projects/responsible-ai-pattern-catalogue>.



Today, numerous organizations are committed to being responsible in the design, development and deployment of AI. However, even with a commitment to responsible AI, it is unclear whether any of these responsible AI efforts really enable responsible AI in practice.<sup>5</sup> According to a recent survey, only 11% of U.S. executives reported fully implementing “fundamental responsible AI capabilities,”<sup>6</sup> and less than half of IT executives in companies with at least U.S.\$100 million in revenue reported that their organization is prepared to invest in responsible AI initiatives.<sup>7</sup> Further, a recent survey of AI startups found that 58% have a set of AI principles, but very few have taken tangible actions—such as firing an employee, dropping training data or saying no to a sale—because of these principles.<sup>8</sup>

Leaders find themselves in a dilemma. On the one hand, they want to “do the right thing” and make responsible AI real in their context. On the other hand, there is a fear that proactive action may result in the kind of “ethics-washing” and “virtue-signaling”<sup>9</sup> that characterized previous corporate responsibility efforts. The following questions currently plague leaders who are designing, building and deploying AI-enabled

systems: How should we operationalize our commitment to responsible AI? And, particularly for leaders not in large corporations, where and how do we start?

So far, part of the response to where to start consists of a confusing array of frameworks, guidelines and government policies.<sup>10</sup> Many large organizations—like Microsoft, Unilever, McKinsey (Quantum Black), IBM, Google and Meta—have a set of (remarkably similar) responsible AI principles describing their proactive commitment. Their corresponding actions vary in form, from technical fixes to AI assurance processes,<sup>11,12</sup> responsible AI standards, and governance and operation models.<sup>13,14</sup> Consequently, many organizations may believe that the way forward is to simply “pick and mix” from some of the principles commonly associated with responsible AI, such as fairness, explainability, accountability and privacy.<sup>15</sup> However, adopting a list of principles is a far cry from operationalizing responsible AI. There is a real risk of principle picking becoming a vehicle for “ethics without teeth,”<sup>16</sup> ultimately undermining its purpose.

The difficulties of operationalizing responsible AI are often called the “principle-to-practice gap.” Schiff and colleagues provide six explanations for its existence:<sup>17</sup> 1) mismatched incentives to embed and operationalize the principles; 2) an underestimation of the complexity of the potential impacts of AI; 3) a plurality of disciplines that need to be involved in shaping the responsible use of AI, which leads to a disciplinary divide; 4) the “many hands” issue,

5 For further discussion around the dearth of evidence regarding whether principles really help, see Mittelstadt, B. “Principles Alone Cannot Guarantee Ethical AI,” *Nature Machine Intelligence* (1:11), 2019, pp. 501-507; Qiang, V., Rhim, J., and Moon, A. “No Such Thing as One-Size-Fits-All in AI Ethics Frameworks: A Comparative Case Study,” *AI & Society* (39:4), 2024, pp. 1975-1994. Despite committing to responsible AI, many technology companies have laid off workers focused on responsible AI in recent times, drawing into question the seriousness of their commitments. See Vynck, G. D. and Oremus, W. “As AI Booms, Tech Firms Are Laying off Their Ethicists,” *Washington Post*, March 30, 2023, available at <https://www.washingtonpost.com/technology/2023/03/30/tech-companies-cut-ai-ethics/>; and Dave, P. “Google Splits Up a Key AI Ethics Watchdog,” *Wired*, 2024, available at <https://www.wired.com/story/google-splits-up-responsible-innovation-ai-team/>.

6 PwC’s 2024 US Responsible AI Survey, PricewaterhouseCoopers, 2024, available at <https://www.pwc.com/us/en/tech-effect/ai-analytics/responsible-ai-survey.html>.

7 Renieris, E. M., Kiron, D., and Mills, S. “Building Robust RAI Programs as Third-Party AI Tools Proliferate,” *MIT Sloan Management Review*, 2023, available at <https://sloanreview.mit.edu/projects/building-robust-rai-programs-as-third-party-ai-tools-proliferate/>.

8 Bessen, J., Impink, S. M., and Seamans, R. “The Cost of Ethical AI Development for AI Startups,” *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022, pp. 92-106.

9 “Ethics-washing” refers to people or organizations who use misleading communication to create the impression of ethical behavior, without substantial application in practice. See Schultz, M. D., Conti, L. G., and Seele, P. “Digital Ethicswashing: A Systematic Review and a Process-Perception-Outcome Framework,” *AI and Ethics*, 5, March 2024, 805-818.

10 See Baxter, K. *Ethical AI Frameworks, Tool Kits, Principles, and Certifications—Oh My!*, Salesforce, September 15, 2022.

11 Responsible AI Tools and Practices, Microsoft, 2024, available at <https://www.microsoft.com/en-us/ai/tools-practices>.

12 *The EU AI Act Has Arrived: How Unilever Is Preparing*, Unilever, 2024, available at <https://www.unilever.com/news/news-search/2024/the-eu-ai-act-has-arrived-how-unilever-is-preparing>.

13 *Microsoft Responsible AI Standard v2 General Requirements*, Microsoft, 2022.

14 *Responsibility: AI Governance Reviews and Operations*, Google, 2024, available at <https://ai.google/responsibility/ai-governance-operations>.

15 Agarwal, S. and Mishra, S. *Responsible AI: Implementing Ethical and Unbiased Algorithms*, Springer, 2021.

16 Resseguier, A. and Rodrigues, R. “Ethics as Attention to Context: Recommendations for the Ethics of Artificial Intelligence,” *Open Research Europe* (1), 2021, Article 27.

17 Schiff, D., Rakova, B., Ayesh, A., Fanti, A. and Lennon, M. “Explaining the Principles to Practices Gap in AI,” *IEEE Technology and Society Magazine* (40:2), 2021, pp. 81-94.

where many different people need to be involved, but the organizational structure does not enable shared responsibility; 5) challenges related to the governance of knowledge in organizational contexts, which can thwart effective translation from principles to practice; and 6) an overabundance of instruments, leading people to become overwhelmed.

The range, diversity and persistence of the challenges associated with operationalizing responsible AI indicate that it is not a technical problem but a systemic one.<sup>18</sup> Indeed, the nonlinear impacts of AI as a technology distinguish it from other technologies, where non-systems approaches may be adequate. Accordingly, the approach of this article does not assume that responsible AI needs only technical solutions. Instead, it takes a cybernetic perspective to what responsible AI practices look like,<sup>19</sup> grounded in three connected orientations: systems thinking, a sociotechnical approach and adaptive change.

First, systems-based approaches to responsible AI include not only technical fixes but also higher-order interventions like addressing incentives or broader organizational considerations, such as the purpose, mindsets or paradigms underlying the system itself.<sup>20</sup> Second, a sociotechnical approach widens the sphere of focus beyond the technology in order to consider it in the broader context of technical, social and

environmental factors.<sup>21</sup> Finally, a perspective that accepts that responsible AI is an adaptive problem embraces the idea that solutions are unknown and progress requires learning.<sup>22</sup> These three orientations are infused into this article's participatory action research approach.

Overall, we distill a five-phase process for how organizations can begin operationalizing responsible AI. In each phase, we outline the problem, activities, challenges and evidence of a shift. Further, we provide key insights from the study and actionable recommendations for organizations that are building or implementing AI-enabled products and wish to operationalize responsible AI in their own context.

## Two Journeys

This is a participatory action research study involving two Australian-founded startups, OrgB and OrgT.

### OrgB, Org T and Why Leaders Should Care About Their Responsible AI Journeys

"OrgB" and "OrgT" are both early-stage startups with Australian founders and are both resource-poor organizations, existing in highly competitive and time-sensitive environments. At the time of writing, both organizations were bootstrapping their operations before seeking their first substantial (seed or Series A) funding and were engaged in building AI-enabled products: OrgT was already building an AI-enabled product, and OrgB had AI in its product roadmap in the near term. In both cases, AI refers to adopting and incorporating large language models (like GPT-3) inside recommender algorithms. Since their operations were not considered "high risk" or "unacceptable risk,"

18 Burton-Jones, A., McLean, E. R., and Monod, E. "Theoretical Perspectives in IS Research: From Variance and Process to Conceptual Latitude and Conceptual Fit," *European Journal of Information Systems* (24:6), 2015, pp. 664-679.

19 Cybernetics is an approach to studying complex systems by looking at the interactions between parts through processes like feedback and communication. See Gould, M., Daniell, K. A., Meares, A. and Bell, G. *Re/Defining Leadership in the 21st Century: The View from Cybernetics*, Australian National University & Menzies Foundation, 2022. The importance of complexity perspectives in information systems is explored in Benbya, H. and McKelvey, B. "Using Coevolutionary and Complexity Theories to Improve IS Alignment: A Multi-Level Approach," *Journal of Information Technology* (21:4), 2006, pp. 284-298.

20 Nabavi and Browne call for a systems-thinking approach to responsible AI and propose a framework called the Five Ps, which we adopt in this article. See Nabavi, E. and Browne, C. "Leverage Zones in Responsible AI: Towards a Systems Thinking Conceptualization," *Humanities and Social Sciences Communications* (10:1), March 2023, Article 82.

21 In information systems, the sociotechnical perspective is considered an important orienting idea (also known as an "axis of cohesion"). It considers the technology as well as those who develop and use it, taking into account the social context. See Sarker, S. at al. "The Sociotechnical Axis of Cohesion for the IS Discipline: Its Historical Legacy and Its Continued Relevance," *MIS Quarterly* (43:3), 2019, pp. 695-720. A cybernetic viewpoint expands this further to include environmental considerations as well. See Bell, G. et al. "Do More Data Equal More Truth? Toward a Cybernetic Approach to Data," *Australian Journal of Social Issues* (56:2), 2021, pp. 213-222.

22 See Figure 2.1 in Heifetz, R., Grashow, A., and Linsky, M. *The Theory Behind the Practice: A Brief Introduction to the Adaptive Leadership Framework*, in *The Practice of Adaptive Leadership: Tools and Tactics for Changing Your Organization and the World*, Harvard Business Press, 2009.

**Table 1: Overview of Two Startup Organizations**

Company Pseudonym	OrgB	OrgT
<b>Region</b>	Australia	Australia
<b>Type of AI</b>	Recommendation algorithm	Recommendation algorithm
<b>Status of AI</b>	AI in product roadmap	AI in current product
<b>Founded</b>	2022	2021
<b>No of co-founders: speciality</b>	4: developer, strategy/fundraising, sustainability (2)	3: developer, strategy/fundraising, legal/fundraising
<b>Stage of funding</b>	Seeking Seed / Series A	Seeking Seed / Series A
<b>Type of business</b>	Sustainability tech	Media Tech
<b>Considered high risk according to EU AI Act</b>	No	No

according to the EU AI Act,<sup>23</sup> they were unlikely to face significant regulatory pressure to undertake responsible AI action. Nevertheless, they both wanted to “be responsible” in their AI use and were unsure where to start (see Table 1).

OrgT is a media-tech startup that strongly believes in ensuring public access to a variety of media sources, highly aligned with target 16.10 of the United Nations’ sustainable development goals (SDGs).<sup>24</sup> OrgT’s product encourages critical thinking by bursting users’ news “bubbles,” recommending a diversity of news perspectives. At the time of writing, their product included a recommender algorithm, leveraging GPT-3, and was publicly available as an AI-enabled Chrome web browser extension (with an app in progress).

OrgB is a sustainability tech startup that supports small and medium enterprises (SMEs) in first identifying which SDGs matter most to them and then providing accessible, actionable recommendations to help SMEs achieve their goals. At the time of data collection, OrgB’s product was a non-AI-enabled app, but an AI-enabled recommender algorithm is in their product roadmap in the near term as well.

While we recognize the challenges in generalizing the findings of this article to larger organizations and flag these challenges below, we nevertheless believe that the insights gained from OrgB and OrgT are interesting and relevant for larger organizations for two reasons. First, following the premise of frugal innovation,<sup>25</sup> if a startup can find a way to operationalize responsible AI, then it is likely that larger, better-resourced organizations can do so as well—and can also adapt the demonstration approach described here to their organizational context. We note some of the places where that adaptation may be necessary. Second, focusing on entrepreneurial organizations building low-to medium-risk AI systems may open new ways of thinking and doing that can be adopted and adapted for riskier systems but, for various reasons,<sup>26</sup> could not have been developed within such contexts.

### Data Collection and Analysis

Data was collected over 2021-2024 through participatory action research,<sup>27</sup> meaning that the primary researcher was an active participant alongside the co-founders in the workshops.

23 The activities of OrgB and OrgT would likely be classified as “limited” or “minimal” risk. See *EU AI Act: First Regulation on Artificial Intelligence*, European Parliament, August 6, 2023, available at <https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>.

24 *The Role of Media: Driving Change towards the SDGs*, Media Development Investment Fund, 2023, available at <https://www.mdif.org/news/role-of-media-driving-change-sdgs>.

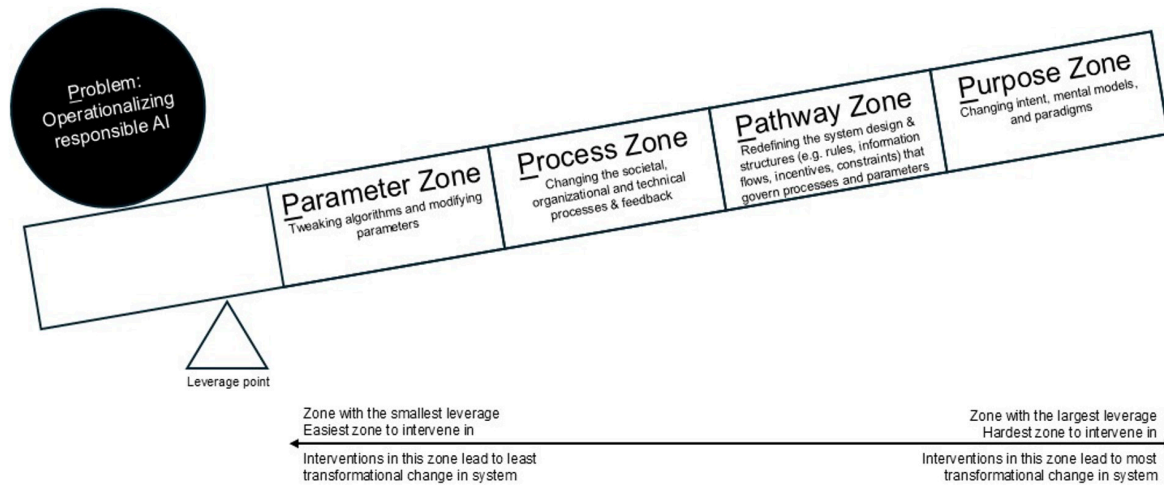
25 Hindocha, C. et al. “Defining Frugal Innovation: A Critical Review,” *BMJ Innovations* (7:4), 2021, pp. 647-656.

26 Barriers to developing solutions in riskier AI system contexts include issues surrounding legacy systems, stuck governance structures, the challenges of large teams and entrenched, unhelpful, cultural practices. See Schiff, D., Rakova, B., Ayeshe, A., Fanti, A. and Lennon, M., op. cit., 2021.

27 Baskerville, R. and Myers, M. D. “Making IS Research Relevant to Practice,” *MIS Quarterly* (28:3), 2004, pp. 329-335.



Figure 1: Overview of the 5Ps Framework<sup>30</sup>



The 5Ps framework: The first 'P' is Problem. In this case, the overall problem is operationalizing responsible AI. At each stage of the 5-phase process, we situate a specific part of the problem within a 'leverage zone'. The remaining 4Ps – Parameter, Process, Pathway, Purpose – describe the 'leverage zones' or places to intervene in the system to address the problem; the zones with largest leverage are hardest to intervene in but have the greatest potential for transformational change of the system.

In the case of OrgT, the three co-founders and the primary researcher worked together in 18 participatory workshops, four group interviews and three one-on-one interviews in 2021-2023. In the case of OrgB, the four co-founders and the primary researcher worked together in six one-on-one interviews and 10 participatory workshops in 2022-2024. (See Appendix 1 for more on the research approach.)

Systems thinking is infused throughout the research and forms the basis for analysis of the data collected. Specifically, the activities that took place were analyzed with a systems thinking framework—the 5P framework—in mind.<sup>28</sup> The 5P framework (Figure 1) identifies places where interventions can be effective in changing the dynamics of a system and was specifically developed “to help improve systems thinking literacy in relation to Responsible AI.”<sup>29</sup>

We thematically analyzed the interactions with OrgT and OrgB on their responsible AI journeys, distilling them into five phases presented in the next section. Each phase consists of a *problem* that emerged (one of the “Ps” in the 5P

framework) and a leverage zone (the other 4Ps)—purpose zone, pathway zone, process zone, parameter zone—within which a response was taken to shift the system. In adopting a systems approach, no problem is ever fully solved due to the ever-changing nature of responsible AI. Nonetheless, steps can be taken in a helpful direction. We also describe activities undertaken to tackle the problem, the artifact(s) that can be of assistance and the challenges faced. Finally, we provide evidence of each phase’s effectiveness in shifting behavior. Insights from conducting the study and recommendations for leaders follow.

## Five Phases to Operationalizing Responsible AI

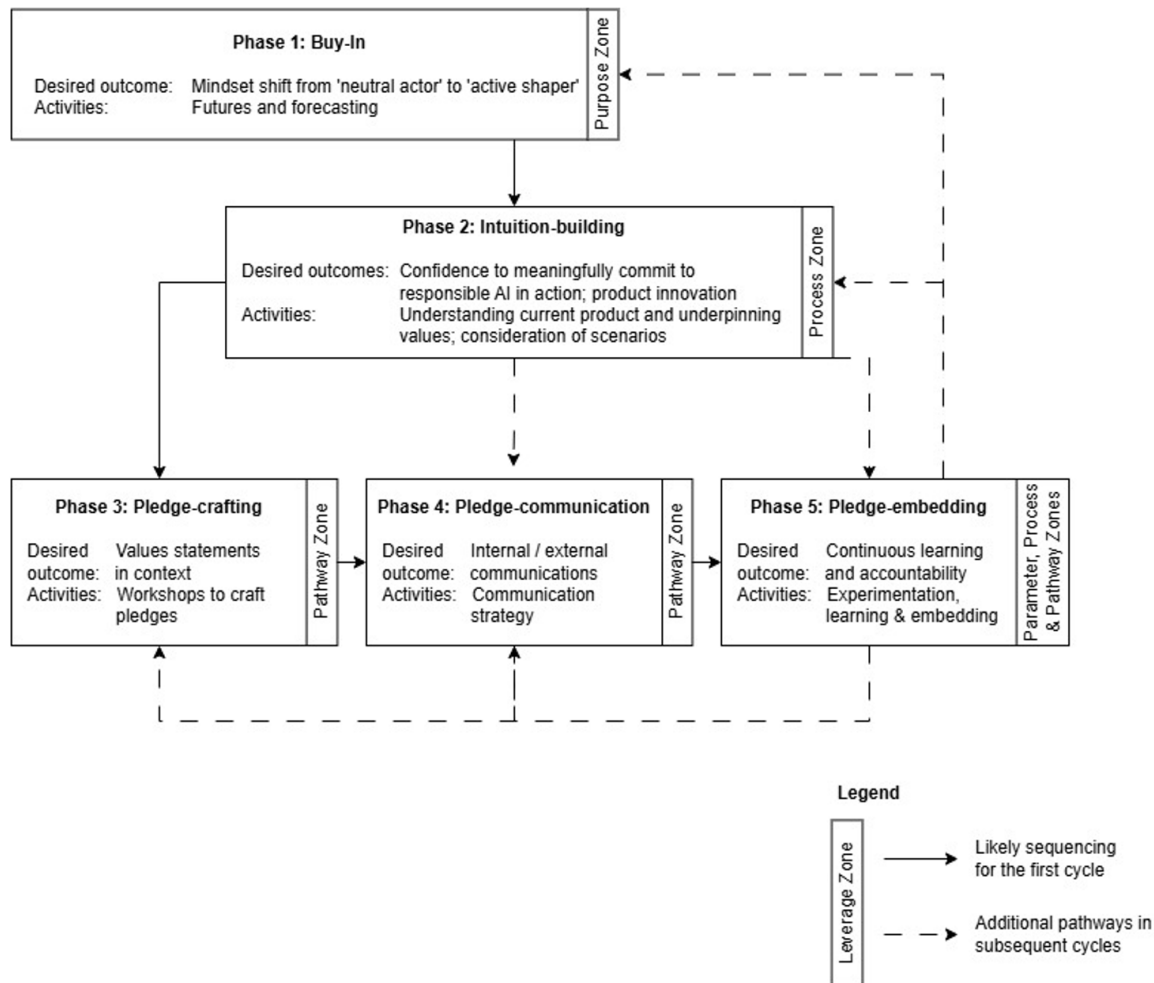
Figure 1 distills a five-phase process for operationalizing responsible AI gathered from the experiences of two organizations. In both cases, the organizations were committed to responsible AI, unsure of where to begin and willing to collaborate with the primary researcher to work out how to start. In the beginning, it was unclear whether or how their individual experiences would relate to each other. While interacting

28 Nabavi, E. and Browne, C. “Leverage Zones in Responsible AI: Towards a Systems Thinking Conceptualization,” *Humanities and Social Sciences Communications* (10:1), 2023, Article 82.

29 Ibid, see p. 4.

30 Ibid, Adapted from Figures 1 and 2.

Figure 2: Overview of the Five-Phase Process for Operationalizing Responsible Use of AI



in the two organizational contexts, common questions and concerns emerged.

For example, what is my role (as an individual) and our role (as an organization) in enacting responsible AI? How do we understand and articulate the values that are important to us? What is our commitment to responsible AI—and how do we communicate it? How do we ensure we walk the talk? Analyzing the interactions through a systems thinking lens assisted us in finding commonalities between the two contexts and shaping the distillation of the phases (Figure 2).

The first two phases focus on generating the right underlying conditions:

- **Phase 1: Buy-in.** This phase involves moving through barriers of skepticism and perceived role neutrality, with organizations accepting their role as active shapers of technology development and use. Futures and forecasting activities unlock this perspective, while constructing future narratives helps make this shift.
- **Phase 2: Intuition-building.** This phase involves building an “intuition” about what responsible use of AI looks like in practice, reducing overwhelming emotions by drawing on decisions already made, and opening future pathways by considering various scenarios. A range of visualization artifacts support this phase.

Three phases associated with pledges follow the foundational phases:

- **Phase 3:** Pledge-crafting. This phase focuses on creating contextualized statements that help turn a commitment to a principle into action. It addresses barriers like fear related to inaction and participation in ethics-washing activities and concerns about the principles' relevance to each organization's day-to-day context.
- **Phase 4:** Pledge-communicating. This phase focuses on how much to operationalize and how much to communicate about the actions taken. It openly addresses fears of backlash for over- or undercommitting to principles that can hinder progress toward operationalizing responsible AI.
- **Phase 5:** Pledge-embedding. This phase can involve many forms of experimentation to continue making pledges real in day-to-day decision-making. It solidifies "walking the talk" on the pledges made.

Although there is an order to these phases, they are not linear; parts of each phase may be unresolved before the next phase begins or revisited over time. The process is also not meant to be exhaustive but rather a path toward recognizing that the work of responsibility is never really done. Learnings from undertaking Phase 5 (pledge-embedding) catalyze new activity cycles.

## Phase 1: Buy-In

**Problem and Leverage Zone:** Phase 1 is focused on ensuring that an organization has adopted and internalized the interactional stance of technology—that is, they believe they are (individually and as an organization) active shapers of technologies and thus of futures.<sup>31</sup> The interactional stance opposes the idea of people and organizations as neutral actors, because such a mindset can lead to a minimization of responsibility, thereby increasing the risk of ethics-washing activities. Phase 1 operates in the purpose zone because it is connected to mindsets and worldviews.

31 Friedman, B. and Hendry, D. G. *Value Sensitive Design: Shaping Technology with Moral Imagination*, MIT Press, 2019.

**Activities Undertaken and Challenges Faced:** The activities undertaken in Phase 1 to facilitate an active-shaper mindset may vary.

For OrgB, this foundational mindset of the active shaper of technology was already present. This phase did not therefore require much attention. For OrgT, on the other hand, their initial mindset was that they believed that the product they were building was "neutral" and that, by extension, they were neutral actors in its development. They were, however, open to exploring this position, which allowed for this phase to occur. For larger organizations, such openness is required from at least the organization's leaders in order to proceed with the five-phase process.

Three core activities were undertaken to shift the perspective from neutral actor to active shaper.<sup>32</sup> First, we brainstormed the major moments, events, products and services that have influenced information consumption from 1960 to today. Through this activity, OrgT began to see itself as part of an evolving ecosystem of interventions that were moving media consumption towards particular futures and began to recognize that these futures were open to influence. Second, we collected signals and drivers relevant to their market context and crafted stories about the future. We discussed which forecast narratives were possible, plausible, probable and/or preferable.<sup>33</sup> Third, we discussed the impact of these forecast narratives on potential OrgT users in 10 years' time. These three activities convinced OrgT that their product is not separate from the market and that futures are created; rather than being passive or neutral observers, they are active shapers with agendas and influence.

**Evidence of a Shift:** By the end of this phase, a change in perspective from neutral actor to active shaper was evident in OrgT. One co-founder told us:

*"We valued considering different potential users, how the futures we had identified*

32 The futures and foresight activities were adapted from Institute for the Future. See *IFTF Foresight Training*, Institute for the Future, available at <https://www.iftf.org/foresightessentials>. Toolkits were found on Observatory of Public Sector Innovation. See also Futures & Foresight, OECD, available at <https://oecd-opsi.org/guide/futures-and-foresight>.

33 See Voros, J. "A Generic Foresight Process Framework," *Foresight* (5:3), 2003, pp. 10-21.



*would impact their day-to-day lives and the role that OrgT plays in their futures. We take our responsibility as entrepreneurs seriously to ensure that we don't make the same mistakes of the past. The exercises we did and the practices that emerged are helping us to avoid drinking our own Kool-Aid. This is particularly important to us and our integrity as an organization—we're in the business of assisting others in identifying bias, so we need to be great at it too!"*

## Phase 2: Intuition-Building

**Problem and Leverage Zone:** At the beginning of our interactions, both OrgT and OrgB were very clear that they wanted to “walk the talk” on responsible AI: “I think most important for me that we actually do what we’re saying,” said an OrgB co-founder. However, given the uncertainty around the concept of responsible AI, managers struggle with applying responsible AI principles without inadvertently succumbing to ethics-washing. Phase 2 was dedicated to establishing and strengthening an “intuition” about what responsible AI principles could look like in their context so that the organization could more confidently commit to responsible AI pledges. Phase 2 operated in the process zone because it focused on creating feedback loops regarding (current and potential) decisions made in product development—and the implications for responsible AI.

**Activities Undertaken and Challenges Faced:** Intuition-building involved a dance between, on the one hand, understanding what was already built and the values underpinning choices already made and, on the other hand, imagining what the product could look like with a commitment to certain principles in place. Practically, this involved considering scenarios in two threads. First, we considered various “common” responsible AI principles—such as transparency, privacy, fairness and accountability—in the communications of well-known or competitor organizations and discussed their relevance to their current product. Second, we considered the explicit or implicit values already existing in their organizational context and how they influenced (or did not influence)

product decisions to date or could influence product decisions in the future.

A detailed understanding of how the product works was necessary to do this. This was a nontrivial exercise because, as is common practice in startup (and other) environments, different co-founders worked on different parts of the product and its delivery, often at the same time and in fast iteration cycles. To assist with this challenge, various visualizations—such as data maps, concept maps and product matrices—served as useful tools to generate collective understanding (also known as boundary objects).<sup>34</sup>

The purpose of this phase was not necessarily to resolve what commitments would be made but rather to build confidence by recognizing that they were already making value-led decisions at times and could imagine other feasible ways of doing so.

**Evidence of a Shift:** An increase in the confidence to take action to embed principles into product decisions can be seen through OrgT’s exploration of their commitment to autonomy and freedom in their product. Data maps that we co-created revealed that their commitment to incorporating a diversity of media sources was at odds with how the product was currently operating. Namely, OrgT had inadvertently decided to limit the number of articles scraped to the top 100 results on Google to economize on the number of tokens (or pieces of words) used for natural language processing by GPT-3. Because OpenAI charges by token,<sup>35</sup> economizing on the number of tokens would, in turn, minimize costs. While there will always be limits on the number of news sources they could use, they recognized that economic drivers had unnecessarily overridden the consideration of other options that would have more adequately upheld their commitment to autonomy and freedom. Consequently, they experimented with ways to achieve a better representation of sources

34 Boundary objects are concepts and things that enable people from different social worlds to have a common conversation. See Star, S. L. “This Is Not a Boundary Object: Reflections on the Origin of a Concept,” *Science, Technology, & Human Values* (35:5), 2010, pp. 601-617.

35 OpenAI charges users of their models by the number of “tokens” used: 1,000 tokens is roughly equivalent to 750 words used for natural language processing. The price per token depends on the model used and its capability. See *Pricing*, OpenAI, 2023, available at <https://openai.com/pricing>.

(upholding their value commitment) while maintaining similar costs.

For OrgB, they strengthened their responsible AI intuition by using ChatGPT. They asked ChatGPT to give suggestions on how to embed each of their draft organizational values into each stage of the product development process and then mapped their progress against each suggestion. The heatmap assisted the OrgB co-founder in questioning the extent to which a certain value was important to OrgB right now. As a result of the matrix, OrgB fast-tracked some activities planned for a later date to solidify their commitment to existing values. They also fundamentally rethought their commitment to a few draft values that they realized were deprioritized in their current decision-making and intentions.

Overall, we see evidence of intuition-building, resulting in more confidence to act in alignment with principles in this phase. In the case of OrgT, it led to product innovation ideas that strengthened the organization's commitment to particular values and increased confidence in their ability to commit to them without succumbing to ethics-washing. In the case of OrgB, it led to the fast-tracking of existing ideas to crystallize commitments, a reconsideration of what values should be chosen in the first place and a reprioritization of what was important.

### Phase 3: Pledge-Crafting

**Problem and Leverage Zone:** With the confidence built in Phase 2 that they could meaningfully “walk the talk,” the next problem to tackle focused on how best to articulate a commitment to responsible AI that positioned the organizations for meaningful, trusted action. This phase sits in the *pathway zone*, as it requires a response that sets up the rules or structures (in this case, pledges) that govern how to do responsible AI in their context.

**Activities Undertaken and Challenges Faced:** The first activity undertaken was to decide to craft pledges instead of simply writing out responsible AI principles. Pledges can be thought of as “promises—commitments to yourself, your team, your customers, various stakeholders and the planet—to see better outcomes in context.”<sup>36</sup> The primary researcher suggested pledges to

replace merely writing a list of responsible AI principles; this was met with enthusiasm from both organizations for two main reasons.

First, writing in a pledge format emphasizes an action orientation, which resonated with both organizations and alleviated some of their concerns about this work being “just fluff that gets put on the website or things that get thrown around when it’s convenient,” in the words of an OrgB co-founder. Second, a pledge clearly shows that the value should be considered in a particular context, narrowing the scope, increasing the perceived do-ability of the pledge and allaying credibility concerns and the risk of losing customer trust in the process.

There are many ways in which a pledge can be formulated. However, pledges are often framed in a way that makes them useful for guiding product decisions. For OrgB, the pledge format included a contextual statement, an aspiration and a demonstration example. Four pledges were crafted and used in the earliest stages of AI system design, with a view of their continued embedding throughout the product development lifecycle. For OrgT, the pledge format included an overarching commitment and a demonstration example. Five pledges were crafted once an AI-enabled product had already been developed; the pledges were used to review what had already been done and to guide future development. Practically, workshops were initiated to draft, redraft and refine the pledges in multiple iterations. For OrgT, all co-founders (and thereby all employees) were present in these workshops. For OrgB, it was refined in between workshop interactions. For larger organizations, consideration would need to be given to who is involved in crafting pledges and what types of feedback loops would be needed for wider organizational input.

It was also recognized that the pledges are open to review and renewal, as products and organizations evolve. For OrgB and OrgT, this review and renewal occurred organically through product development conversations; however, for larger organizations, formalizing when and how the renewal will take place may be needed. Example pledges from OrgB and OrgT are captured in Table 2.

**Evidence of a shift:** By the end of this phase, both organizations felt they had captured

<sup>36</sup> *Pledge Works*, ResponsibleTech.Work, March 1, 2022, available at <https://responsibletech.work/tools/development/pledge-works>.

**Table 2: Example Pledges**

Company Pseudonym	Value	Pledge
OrgB	Empower & Include	[contextual statement] We believe that sustainability actions are every SME's business. We exist to empower all SMEs to take these sustainability actions. [aspiration] Empowerment goes beyond 'just helping' and inclusion goes beyond 'just accessibility'. [demonstration example] For us, this looks like designing products that are simple and accessible, using language that suspends all judgement and meets SMEs where they're at, identifying and counteracting biases in our approach where we can and fostering a wider collaborative ecosystem of mission-aligned partners.
OrgT	Autonomy & Freedom	[general commitment] We pledge to ensure that OrgT is an enabler of autonomy and freedom. [demonstration example] For us, this means making sure that the OrgT platform contributes to your ability to engage in critical thinking and come to your own conclusions by giving access to a plurality of perspectives, tailored to you.

responsibility commitments that they could meaningfully adhere to, alleviating the risk of eroding customer trust. An OrgT co-founder told us:

*"I think people knowing that we've critically thought about it and we've got a method to mitigate ... I think the audience we're targeting will be like 'Cool. Clearly on it.' We want the audience to look at us and go 'Cool. They're on it. They're clearly deep in this world [of Responsible AI]. We trust [OrgT]. We trust it [OrgT's product].'"*

#### Phase 4: Pledge-Communicating

**Problem and Leverage Zone:** With pledges articulated from Phase 3, both organizations turned to the problem of whether and how to communicate these pledges externally. This phase operates in the pathway zone, as it is concerned with information flows and access to information.

**Activities Undertaken and Challenges Faced:** The primary activity undertaken in this phase was an open discussion about how best to communicate the pledges externally. In both organizations, all co-founders had been involved in crafting the pledges, and there were no further employees to consider. For larger organizations, however, how best to deal with the internal communication of pledges before communicating them externally may also need to be considered.

For OrgB, their brand was already highly aligned with taking responsible action; thus,

leading by example—in both taking action and externally communicating their actions—was a straightforward decision. But in the case of OrgT, how to communicate the pledges was an extended, considered conversation. We anticipate this may be the case for larger organizations as well.

We discussed navigating the thresholds between doing too much and doing too little while communicating too much and too little across different stakeholder groups (see the Key Insights section for further detail).

**Evidence of a Shift:** Both organizations decided to integrate the pledges into their websites. OrgB also integrated them into funding pitch decks. For OrgT, they decided to include a separate file in their data room for prospective funders, believing that funders' reactions to this could help weed out non-value-aligned parties.

#### Phase 5: Pledge-Embedding

**Problem and Leverage Zone:** With pledges articulated and communicated—and confidence that they could meaningfully act—the central problem of this phase was how to continually experiment and embed the pledges in day-to-day practice. This phase can operate in any of the leverage zones but is most likely to involve the parameter zone (e.g., tweaking numbers or criteria within algorithms), the process zone (e.g., creating or modifying feedback loops) and the pathway zone (e.g., changing information flows, incentive structures or constraints). Learnings

from the embedding stage catalyze new cycles of activity, which may take many paths, as per Figure 2.

**Activities Undertaken and Challenges Faced:** The activities undertaken in this phase are highly contextually dependent. Interestingly, both organizations used generative AI to experiment with holding themselves accountable for the pledges they had recently made. In both cases, the developer co-founder drove experimentation: one focused on how to hold themselves accountable for the pledges during the coding process; the other focused on how to hold themselves accountable for the pledges during product development.

Challenges faced in this phase included keeping the pledges front and center in the AI development process while undertaking their everyday activities. For example, developers in both organizations lamented the difficulties of zooming out to consider the principles and then zooming in to adjust the code.

**Evidence of a Shift:** Having a set of responsible AI principles that sit idle and are never operationalized is common. In contrast, we see evidence of experimentation to embed the pledges from both organizations, instigated at their own volition and outside of the researcher interactions.

The OrgT developer co-founder undertook three experiments to integrate thinking about the responsible AI pledges while coding. The initial experiment was analog: handwritten pledges on a piece of paper were placed alongside the computer during coding. Even this simple approach yielded additional reflections on the relevance of the pledges to the coding decision-making process. The developer co-founder told us: “To be honest, every time I looked at it [the pledges piece of paper] I found something in the code where I was like, ‘Actually, you know what, this is relevant!’ I thought less of the code would be impacted [by the pledges] than it was.”

The second experiment involved the developer environment GitHub Copilot.<sup>37</sup> Pledges were added at the top of the coding environment as a comment, and comments were used throughout to connect the pledges to the decisions made. Over time, GitHub Copilot began suggesting

comments. Although the relevance of these suggested comments was not particularly high, it may improve over time. As the developer co-founder explained: “My hope was that because I used [GitHub] Copilot to code that it would recognize these [pledges] as part of the thing I’m trying to do and might start to suggest where those ethics are relevant.”

The third experiment was undertaken with the launch of a new developer environment, Cursor,<sup>38</sup> which integrates a chat function pane so developers can “chat with their code” around technical difficulties and bug fixing. OrgT’s developer co-founder co-opted this functionality for pledge accountability, using it to specify that they wanted code written with specific pledges in mind. It was also asked to label instances where violations of particular pledges might occur in order to serve as warning labels for others reusing the code.

OrgB experimented with ensuring that their pledges would be included in their day-to-day decision-making by focusing on the product development process. The experiments used ChatGPT to create a matrix of product considerations aligned to their pledges. Following a sense-check of the matrix for appropriateness, the co-founder created a heatmap, shading different parts of the matrix according to whether the suggested action was already incorporated, could be better incorporated, would happen at a later (known/scheduled) time, was intended for a later (undefined) time or was not included in the product roadmap at all. This heat map served as a conversation tool to help the co-founder understand what values were already embedded in their decision-making (contributing to Phase 2, “building an intuition”) and—when repeated at a later time—the extent to which the pledges were being used in their latest practices (Phase 5). The co-founder said:

*“Overall, we’ve got more green [action already incorporated] now, compared to the earlier check, which is great. I’d say this is a combination of us focusing and consolidating the values and all the work we’ve done in the interim. Most of the grey [not yet included in the product] items*

37 *GitHub Copilot—Your AI Pair Programmer*, GitHub, available at <https://github.com/features/copilot>.

38 *Cursor—The AI-First Code Editor*, Cursor, available at <https://cursor.sh>.



*from earlier were in the values that we dropped—this is another proof point that these values weren't really resonating with us/the product."*

## Key Insights

In this section, we share some key insights from engaging with the journeys of OrgB and OrgT.

### The Approach Taken to Operationalize Responsible AI Is as Important as the Actions

In this study, the systems-based approach, alongside working in collaboration with the organizations through action research, was important to reduce some of the known barriers to operationalizing responsible AI.<sup>39</sup> Some of these barriers were also reduced due to the small, nonhierarchical nature of startups; for larger organizations, efforts to address barriers regarding the disciplinary divide and the many-hands issue may need more attention. In contrast, the barrier associated with the knowledge governance regarding the possible implications of AI was amplified by the general lack of governance structures in the startup context. Accordingly, larger organizations may be able to access more organizational pathways to address this barrier (see Table 3).

### Investing in Building Responsible AI Muscles in the Early Phases and in Small Projects Is a Promising Approach

Our experiences challenge a commonly held concern that it is a waste of time and resources to invest in responsible AI when projects are small or early in the development process—and even more so in the startup context, as the likelihood of organization failure is high. Instead, we observed the fruits of the responsibility muscles fostered in these environments and suggest that there is value to this approach.

For OrgB, AI was still not embedded in their product. However, the responsibility muscles built through the process were seen in everyday decision-making, such as in how they performed quality assurance on suppliers

and in the products they recommended through their platform (and products that will be recommended by AI in the future). "The algorithm is me at the moment," said one OrgB co-founder, "and I hadn't realized that I really do need to sense check everything [against the pledges]." Alignment with the pledges also surfaced as part of OrgB's funder-vetting processes.

For OrgT, their activities were "on hold" 15 months after the pledge-making process, due to challenges in raising sufficient capital. However, the OrgT co-founders are now working together on another AI build project with a paying customer, and the responsibility muscles built will transfer to their new project. As an OrgT developer co-founder explained:

*"I wanted to understand how the data was gathered [in the new project] and make sure we have some information about privacy. I don't think I would have thought about that stuff much before [the work we did together]. There's like a little muscle in my brain now that twinges when people start talking about big data sets. The ethics principles, I think it's almost more important now [with the current project]. I don't want to put more crap into the world."*

### Doing Too Much vs. Too Little and Communicating Too Much vs. Too Little Can Hamstring Organizations

Navigating tensions between doing and communicating too much or too little is important for operationalizing responsible AI. As we learned with OrgT, communicating externally raises a range of concerns around ethics-washing, virtue-signaling, commercial viability and customer trust. Left unresolved, it can leave organizations hamstrung and paralyzed. Since external communications can often serve as an accountability mechanism for sustained responsible AI action over time, fostering meaningful external communication is a responsible AI action in itself.

For OrgT, there were many concerns. They did not want to "talk too much": They were inclined to say less and overdeliver rather than overpromise and be accused of not walking the talk. Equally, they recognized that if they did not

<sup>39</sup> See Schiff, D., Rakova, B., Ayesha, A., Fanti, A. and Lennon, M., op. cit., 2021.



**Table 3: Assessment of Proposed Process Against Explanations for Principle-to-Practice Gap in Responsible AI**

Explanations for the principle-to-practice gap in responsible AI	Extent to which this explanation is addressed in the proposed process
Misalignment of incentives	Addressed through 1) a starting condition of “wanting to be responsible” and 2) product innovation benefits from engaging in the process.
Underestimation of the complexity of AI impacts	Addressed through the systems-based approach taken in this study.
A disciplinary divide	Addressed. Problem is mitigated from the outset due to the smaller organization size in the startup context.
The “many hands” issue—organizational structures that do not enable shared responsibility	Addressed. Problem is mitigated from the outset due to the smaller organization size and nonhierarchical structure in the startup context.
Governance of knowledge about possible implications of AI	Partially addressed. Problem is exacerbated by the startup context, which is less likely to have existing policies and processes that can be leveraged.
Overabundance of tools	Addressed by focusing on pledge-making and by the researcher accompanying the organization throughout the process.

communicate enough about their good work, they might not fully exercise their responsibility of leading the field or might miss opportunities to be held accountable by others.

Further, there was much discussion about potential backlash from future funders if they “do too much” and communicate openly about it. An OrgT co-founder questioned: “Do they [funders] see that [ethical behavior] as a bug or a feature? Do they see it as, ‘These guys are slowing down, let’s just break things and run with it.’ Or they’re like, ‘Okay, good, this is sexy, ethics is sexy now.’”

They worried that a perceived excess of responsible activity might be interpreted as undermining their commerciality or poor use of scarce resources. Yet “doing too little” could negatively influence whether their user base sees them as trustworthy. Figure 3 depicts the different thresholds that need to be navigated.

## Recommendations for Leadership

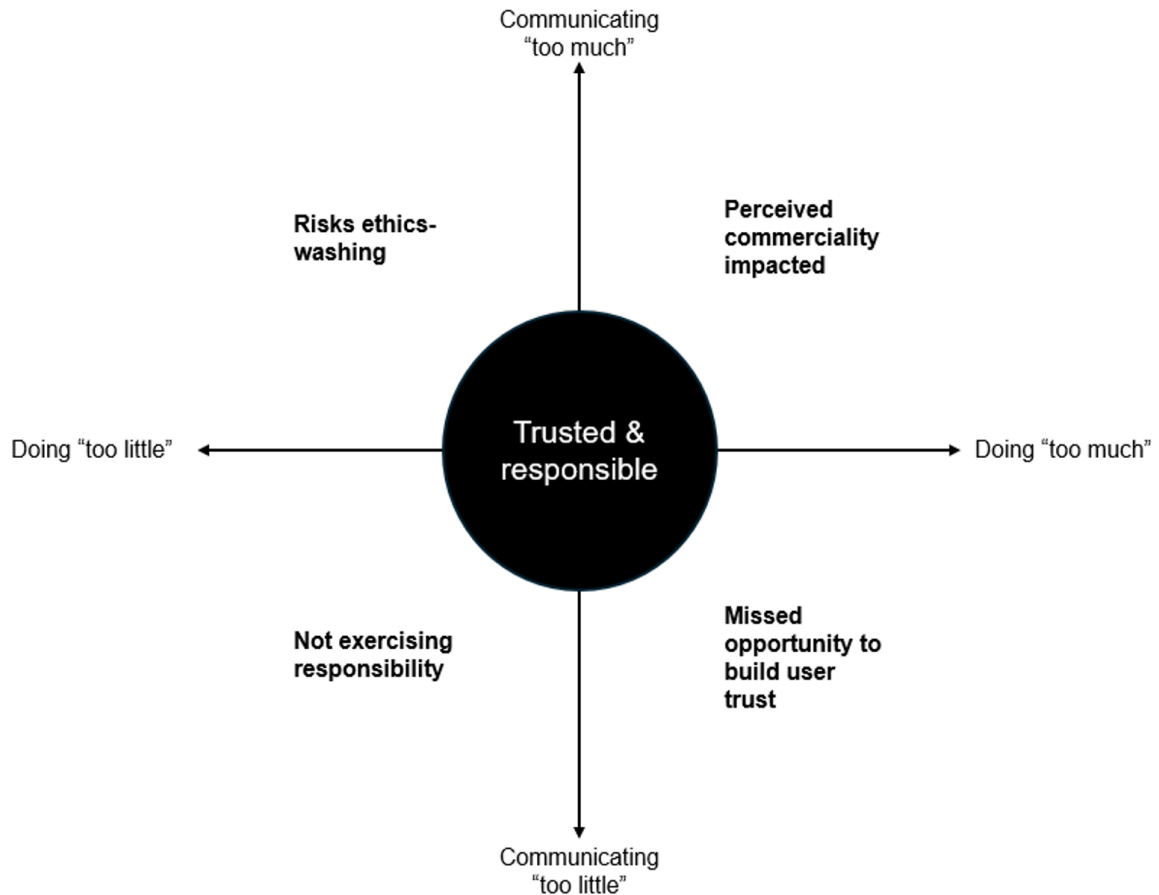
Based on an analysis of the activities that occurred with OrgB and OrgT, Table 4 provides a roadmap to support executives beginning to operationalize responsible AI use. It contains an overview of the five-phase process, with the corresponding problem, leverage zone and

challenges faced as well as the activities and artifacts that can be of assistance. It also gives examples of some outcomes achieved through the process, noting that these are illustrative and not exhaustive. Additional considerations for larger organizations are also included where appropriate. Five recommendations follow.

### R1: Approach Operationalizing Responsible AI from a Systems Perspective

The primary premise of this article is that operationalizing responsible AI is not a siloed technology problem, but a systems problem.<sup>40</sup> Accordingly, much consideration needs to be given to aspects outside the technology itself—specifically, the mindsets of those within the organization as well as the cultures of iteration and experimentation. Leaders must acknowledge that responsible AI is not a checklist activity but a set of practices to be built over time, subject to ongoing renewal and reassessment. Without a systems perspective, organizations risk pouring

40 This follows the proposal for improving outcomes toward responsible AI, via systems-based approaches made by Nabavi, E. and Browne, C., op. cit., March 2023. It also aligns with the cybernetic approach of seeing technology as part of a system. See Daniell, K. et al. *Response to the Inquiry into the Use of Generative Artificial Intelligence in the Australian Education System*, Western Sydney University, 2023.

**Figure 3: Communication and Action Thresholds to Navigate**

time, energy and other resources into efforts that may not stick—or, worse, activities that may result in ethics-washing. Thoughtfully crafted pledges assist in this and arguably enable further action orientation and contextualization beyond merely stating principles, but such pledges nevertheless cannot operate alone. The five-phase process demonstrates how activities can happen at different system scales to ensure the operationalization of responsible AI.

## **R2: Cultivate a Mindset That Your Organization Has a Role in Advancing Responsible AI (and Don't Start Without It!)**

Ensuring that your organization owns its role as an active shaper of the technology (and not as a neutral or passive party) is a critical first step. As described in Phase 1, futures


and forecasting activities can be employed to nurture an understanding that human choice is involved in designing and deploying AI systems. Embodying the interactive stance of technology—that humans shape technologies, which, in turn, shape humans—is a must before embarking on the rest of the process.<sup>41</sup> Without an embodiment of a larger “why,” it is likely that momentum will be difficult to sustain, resulting in half-executed programs, wasted resources and lost opportunities to capitalize on efforts to be responsible. In larger organizations, the organization’s position may be more fractured; nevertheless, at the very least, a clear position on the organization’s mindset via its leaders is needed before proceeding. Otherwise, there is a real risk of promoting ethics-washing activities

<sup>41</sup> Friedman, B. and Hendry, D. G. *Value Sensitive Design: Shaping Technology with Moral Imagination*, MIT Press, 2019.

**Table 4: Roadmap for Navigating How to Get Started with Operationalizing Responsible AI**

	<b>Phase 1: Buy-in</b>	<b>Phase 2: Intuition-building</b>	<b>Phase 3: Pledge-crafting</b>	<b>Phase 4: Pledge-communicating</b>	<b>Phase 5: Pledge-embedding</b>
Problem	Interactive stance of technology is not internalized, putting any efforts at risk of ethics-washing.	Lack of confidence regarding what responsible AI could look like in practice.	How best to articulate responsible AI commitments.	Whether and how best to communicate responsible AI commitments.	How to make the pledges real in day-to-day decision-making.
Leverage Zone(s)	Purpose zone	Process zone	Pathway zone	Pathway zone	Parameter, Process and Pathway zones
Activities undertaken	Futures and forecasting	<ul style="list-style-type: none"> <li>• Understanding current product</li> <li>• Understanding values underpinning product decisions already made</li> <li>Consideration of scenarios</li> </ul>	Workshops to draft, redraft, and refine pledges	Communication strategy	Experimentation, learning, and embedding (highly contextually dependent)
Challenges faced	<ul style="list-style-type: none"> <li>• Skepticism</li> <li>• Perceived role neutrality</li> </ul>	<ul style="list-style-type: none"> <li>• Fear of ethics-washing</li> <li>• Overwhelm</li> <li>Paralysis</li> </ul>	<ul style="list-style-type: none"> <li>• Inaction</li> <li>• Irrelevance</li> </ul>	<ul style="list-style-type: none"> <li>• Fear of ethics-washing</li> <li>• Potential backlash from funders and other stakeholders</li> <li>• Concerns over the erosion of customer trust</li> </ul>	<ul style="list-style-type: none"> <li>• Stasis</li> <li>• Backlash</li> <li>• Threats to ongoing integrity and trust</li> <li>• Time and resource prioritization</li> </ul>
Assisting artifacts	<ul style="list-style-type: none"> <li>• Constructed narratives about possible futures</li> </ul>	Visualizations (data maps, concept maps, etc.)	Pledge statements	Communication artifacts	Various e.g. coding prompts, decision frameworks, code library documentation
Outcomes	<ul style="list-style-type: none"> <li>• Mindset shift from 'neutral actor' to 'active shaper'</li> </ul>	<ul style="list-style-type: none"> <li>• Confidence they could meaningfully commit to responsible AI action</li> <li>• Product innovation ideas</li> <li>• Values (re) consideration</li> <li>• Product feature (re)prioritization</li> </ul>	<ul style="list-style-type: none"> <li>• Values statements in context, i.e. pledges</li> </ul>	<ul style="list-style-type: none"> <li>• Internal/external accountability</li> </ul>	<ul style="list-style-type: none"> <li>• Continuous learning and accountability</li> </ul>

**Table 4: Roadmap for Navigating How to Get Started with Operationalizing Responsible AI (Continuation)**

	Phase 1: Buy-in	Phase 2: Intuition-building	Phase 3: Pledge-crafting	Phase 4: Pledge-communicating	Phase 5: Pledge-embedding
Outcomes	<ul style="list-style-type: none"> <li>• Mindset shift from 'neutral actor' to 'active shaper'</li> </ul>	<ul style="list-style-type: none"> <li>• Confidence they could meaningfully commit to responsible AI action</li> <li>• Product innovation ideas</li> <li>• Values (re) consideration</li> <li>• Product feature (re)prioritization</li> </ul>	<ul style="list-style-type: none"> <li>• Values statements in context, i.e. pledges</li> </ul>	<ul style="list-style-type: none"> <li>• Internal/external accountability</li> </ul>	<ul style="list-style-type: none"> <li>• Continuous learning and accountability</li> </ul>
Additional considerations for larger organizations	<ul style="list-style-type: none"> <li>• Fractured views within the organization regarding the internalization of interactive stance are more likely.</li> <li>• Leadership must be on board at a minimum before proceeding.</li> </ul>	<ul style="list-style-type: none"> <li>• Will need to consider which parts of the organization need confidence built.</li> </ul>	<ul style="list-style-type: none"> <li>• Consider who needs to be involved in pledge-creation and in feedback loops within the organization.</li> </ul>	<ul style="list-style-type: none"> <li>• May also need to consider how best to communicate responsible AI commitments internally before external communications.</li> </ul>	<ul style="list-style-type: none"> <li>• A culture of experimentation, reflection, and dialogue is needed.</li> </ul>
					

through the process, undermining its ultimate purpose.

### R3: Craft Responsible AI Pledges Instead of Just Stating Principles

Organizations should refrain from the abstracted approach of stating principles used by many large corporate actors and, instead, embrace pledges. Pledges can take different formats, but they must share one thing in common: expressing a principle in a contextualized, action-oriented way. In the cases

of OrgT and OrgB, we used the following format: “We pledge.... For us this means...” (Table 2).

Leaders of larger organizations can also consider cascading pledges with increasing granularity, where pledges at different scales—for example, department, geography, team, individual, etc.—contribute to the larger organizational pledge. Similarly, pledges made to different stakeholder groups beyond just the customer could also be considered.

### R4: Encourage a Culture of Experimentation and Dialogue to Make

## Sense of Operationalizing Responsible AI Over Time

Embedding pledges goes hand in hand with cultures of learning, dialogue and experimentation. Creating visualizations is one way in which this culture is fostered, enabling a shared understanding of product choices and improving organizational cohesion and team trust. Given that an experimental organizational culture may not be the norm for all organizations, particularly larger organizations, such activities may need to be incentivized. As leaders design processes for operationalizing the responsible use of AI, it will be important to honor spaces for experimentation, reflection and dialogue. Without such spaces, those committed to embedding responsible AI may become disgruntled, demotivated or frustrated by obstacles. There may also be missed opportunities for learnings between team members that could enable the operationalization of responsible AI in different contexts.

## R5: Foster Relationships with Entrepreneurial Ecosystems to Learn from Their Responsible AI Journeys

We encourage organizations of all sizes to foster relationships with startups like OrgB and OrgT so that they can learn from their experimentation with operationalizing responsible AI. For startups, this may mean finding communities of practice within the entrepreneurial ecosystem where demonstration cases of responsible AI are shared and discussed. For larger organizations, we encourage a widening of perspectives to learn not only from competitors or other similarly sized organizations but also from other organizations that have the scope to experiment differently with responsible AI practices.

## Concluding Comments

Achieving the potential of artificial intelligence for societal impact starts with acting responsibly when developing and using AI. Despite a strong need for guidance regarding the operationalization of responsible AI, few studies have focused on how to get started practically. This article fills this void by taking a systems perspective and contributing to a demonstration

process involving five phases—buy-in, intuition-building, pledge-crafting, pledge-communicating and pledge-embedding. Although we believe that these findings from early-stage startups can have broader applicability to other, more mature organizational contexts, future research would be needed to verify this.

Based on this study, we conclude that the systems approach to operationalizing responsible AI is an important factor in addressing known barriers. We debunk a commonly held view that responsible AI is only an activity for large organizations and projects, instead showing how focusing on small and early projects can have ongoing benefits. To assist organizations, we also present a thinking tool for navigating the challenge of doing too much or too little responsible AI and communicating too much or too little about it.

To begin, we recommend that organizations embrace a systems approach and start the process only if they possess the mindset that they have an active role to play. With these foundational conditions in place, we recommend that organizations embrace pledges over principles, cultivate a culture of experimentation and dialogue, and foster relationships with startups that may be able to innovate in ways they cannot.

This article is important to responsible AI practice because it demonstrates how to operationalize responsible AI beyond compliance with regulations—namely, by taking a systems approach to creating, communicating and embedding responsibility pledges. It is hoped that with responsible AI operationalized in more organizations, we can ensure the potential of AI for societal benefit.

## Appendix 1: Research Methodology

Data collection was conducted under Australian National University ethics protocol 2022/051, consisting of 18 participatory workshops, four group interviews and three one-on-one interviews with OrgT over 2021–2023, as well as six one-on-one interviews and 10 participatory workshops with Org B from 2022 to 2024. For OrgT, the initial impetus for the work was to assist the organizations with



systems thinking and help them understand its relevance to their product development. This then turned into working with them on their concerns about being responsible in practice. For OrgB, the work began with the goal of reconstituting their organizational values. OrgB had a strong sense of the importance of responsible AI use from the beginning and was keen to engage in this topic in preparation for the use of AI in future products. They believed that the best way to do this would be to crystallize their organizational values and that these would then flow into guiding responsible AI use.

The collected data was recorded and transcribed and then thematically analyzed using NVivo. Field notes following each interaction—as well as the capture of documents, emails, Slack messages, diagrams and blog posts created during our interactions—were also part of the data collection.

The research design was undertaken as participatory action research: The primary researcher was an active participant in the workshops, providing her own input to discussions—in many cases, as “rational myths” to be tested. The rational myths tested in this research included 1) a reflective cybernetic approach (where the researcher and co-founders saw themselves as active participants in shaping the system) as an important grounding for motivational commitment to meaningful, actionable responsible AI; 2) that futures and foresighting can be used as an on-ramp to generating this cybernetic awareness; 3) that responsible AI can be seen as an unfolding, iterative journey of activities; and 4) that reducing the principles-to-practice gap can be achieved by crafting pledges.

pursuing her Ph.D. at Australian National University’s School of Cybernetics, she holds a master’s in applied cybernetics from ANU and a master’s in international management (CEMS MIM) from the University of Sydney, HEC Paris and Copenhagen Business School. She is a Fellow of the International Humanistic Management Association and an Acumen Global Fellow.

### **Katherine Daniell**

Professor Katherine Daniell (Katherine.Daniell@anu.edu.au) is the interim director of ANU’s School of Cybernetics and a John Monash Scholar. With a background in engineering, arts and public policy, she specializes in collaborative approaches to sustainable development and cybernetic systems. Katherine is an expert on how technologies like AI shape society and on participatory governance methods. She has published over 100 academic works, including four books, and has extensive experience in cross-sector research collaboration. Her contributions have earned multiple honors, including being named a French Chevalier in the Ordre National du Mérite.

## **About the Authors**

### **Lorenn Ruster**

Lorenn Ruster (Lorenn.Ruster@anu.edu.au) is a transdisciplinary scholar advancing responsible AI at the intersection of academia and practice. Building on over a decade of professional experience in systems thinking, community-led technology, strategy and governance, Lorenn applies cybernetic, participatory and sociotechnical perspectives to enable responsible AI implementation in organizations. Currently