Principled Practice

A Playbook for Operationalizing Responsible AI

September 2025

Authored by
Beba Cibralic, *Associate Fellow, Leverhulme Centre for the Future of Intelligence, University of Cambridge*
Miranda Bogen*, Director, CDT AI Governance Lab*
Kevin Bankston*, Senior Advisor on AI Governance, CDT AI Governance Lab*
Ruchika Joshi, *AI Governance Fellow, CDT AI Governance Lab*

Introduction

For many organizations, rapid advancements in machine learning has outpaced the creation of robust governance mechanisms. This problem became particularly acute in late 2022 when the popularization of large language models (LLMs) produced a slew of new governance challenges and put a spotlight on existing gaps. As artificial intelligence (AI) diffuses across the economy, an increasing number of organizations — from large to small, from public to private, and across a range of sectors, from finance and technology to education and healthcare — are trying to address both legacy and emerging governance problems. Previously, the work of developing responsible AI was concentrated in a handful of technology companies; today, it is relevant to all organizations using AI in order to prevent clear harms as well as foster consumer trust.

"Responsible AI" (RAI) is an umbrella concept that typically refers to an array of principles, practices, and standards that help ensure AI technologies and products are developed and used safely, ethically, and in line with societal expectations. Responsible AI practitioners are those individuals who work toward these goals, and can range from computer scientists to ethicists, UX researchers, legal professionals, and beyond.

Although attention to responsible AI has ballooned and shape-shifted over the years, consistent practitioner work in this domain contributed to a small-but-growing well of community knowledge — practices, tools, and lessons learned — that can be drawn from by both new practitioners seeking to build effective RAI practices as well as by seasoned ones aiming to enhance their own systems.[1] To capture these learnings from the field, the Center for Democracy & Technology (CDT) brought together over 30 practitioners in the summer of 2024 for a workshop and interviews; the group included socio-technical researchers, designers, policy experts, technical leaders, and compliance and legal personnel, with experience working on efforts related to responsible AI, AI ethics, and AI safety.[2] Practitioners came from a range of organizations across industry, civil society, and government, and their insights revealed **five building blocks** — conceptual and technical infrastructure — needed to operationalize responsible AI. While many practitioners developed their expertise in the technology sector, we expect the contours of their experiences to be broadly transferrable to other domains, and to be illuminating for professionals who are confronting similar issues at an earlier stage of organizational maturity. Those five building blocks (the five "P"s), are:

1. **People: Empower your experts**
   Responsible AI goals are best served by multidisciplinary teams that contain varied domain, technical, and social expertise. Rather than seeking "unicorn" hires with all dimensions of expertise, organizations should build interdisciplinary teams, ensure inclusive hiring practices, and strategically decide where RAI work is housed — i.e., whether it is centralized, distributed, or a hybrid. Embedding RAI into the organizational fabric and ensuring practitioners are sufficiently supported and influential is critical to developing stable team structures and fostering strong engagement among internal and external stakeholders.

2. **Priorities: Thoughtfully triage work**
   For responsible AI practices to be implemented effectively, teams need to clearly define the scope of this work, which can be anchored in both regulatory obligations and ethical commitments. Teams will need to prioritize across factors like risk severity, stakeholder concerns, internal capacity, and long-term impact. As technological and business pressures evolve, ensuring strategic alignment with leadership, organizational culture, and team incentives is crucial to sustaining investment in responsible practices over time.

---

[1] This work builds on previous research on the complex dynamics of operationalizing responsible AI and ethics at technology companies. See e.g. Metcalf, J., Moss, E., & boyd, d. (2019). Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics. Social Research: An International Quarterly. Johns Hopkins University Press, https://datasociety.net/wp-content/uploads/2019/09/Owning-Ethics-PDF-version-2.pdf [https://perma.cc/B9M2-ZETR]; Rakova, B., Yang, J., Cramer, H., and Chowdhury, R (2021). Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for Shaping Organizational Practices. Proceedings of the 24th ACM Conference on Computer-Supported Cooperative Work and Social Computing, https://arxiv.org/abs/2006.12358 [https://perma.cc/2TGT-V84X].

[2] With respect to methodology, we used informal research methods for developing this report. Both the workshop and interviews were semi-structured and unrecorded. Notetakers were present in all sessions. Notes from the sessions were used to inform the report's findings as well as for guideposts on which topics to analyze using secondary sources. This work builds on a foundation of insightful research on responsible AI; these can be found in our citations. Many of the practitioners we consulted came from the technology sector.

3. **Processes: Establish structures for governance**
Organizations need structured governance mechanisms that move beyond ad-hoc efforts to tackle emerging issues posed in the development or adoption of AI. These include standardized risk management approaches, clear internal decision-making guidance, and checks and balances to align incentives across disparate business functions. Processes should layer formal methods (e.g., audits, review checkpoints) with informal ones (e.g., ethical norms, internal culture) to support consistency and institutional memory required for effective AI governance.

4. **Platforms: Invest in responsibility infrastructure**
To scale responsible practices, organizations will be well-served by investing in foundational technical and procedural infrastructure, including centralized documentation management systems, AI evaluation tools, off-the-shelf mitigation methods for common harms and failure modes, and post-deployment monitoring platforms. Shared taxonomies and consistent definitions can support cross-team alignment, while functional documentation systems make responsible AI work internally discoverable, accessible, and actionable. Infrastructure that balances automation with the need for human oversight is particularly crucial for navigating high-stakes contexts.

5. **Progress: Track efforts holistically**
Sustaining support for and improving responsible AI practices requires teams to diligently measure and communicate the impact of related efforts. Tailored metrics and indicators can be used to help justify resources and promote internal accountability. Organizational and topical maturity models can also guide incremental improvement and institutionalization of responsible practices; meaningful transparency initiatives can help foster stakeholder trust and democratic engagement in AI governance.

In discussing each building block, we explore practical and procedural questions responsible AI practitioners will need to navigate as well as challenges they'll likely face in building, adapting, or scaling up responsible AI efforts. Although there is no one-size-fits all approach and the “right” answers will differ for each organization, this report aims to articulate those tradeoffs and their implications so that practitioners can make informed decisions. Others in the field have already mapped out values that drive responsible AI.[3] Building on this work, we help translate these into concrete, actionable organizational considerations.

Throughout this report, we include illustrative, deidentified reflections that responsible AI practitioners shared during our workshops and conversations.

[3] See, for examples: Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. The Journal of Strategic Information Systems. https://www.sciencedirect.com/science/article/pii/S0963868724000672 [perma.cc/DYA3-9YYT]

People: Empower Your Experts

Before any responsible AI work can be done, people with the needed skill sets should be identified and placed in a part of the organization where they can do their work. Responsible AI work is most effectively conducted by a multidisciplinary team of practitioners who can drive nuanced initiatives and ensure that related work meets high quality standards; informed senior leadership to sponsor and advocate for these efforts; and a strong organizational and team structure to ensure the work can take root and flourish. To achieve this, institutions will need to make important design choices in how they organize relevant teams and leadership structures, including which kinds of expertise are valued, whether responsible AI efforts are centralized or diffused, and which leaders oversee the responsible AI portfolio. However, before that can happen, some person or team within the organization needs to recognize responsible AI development and deployment as a priority and promote that priority to the rest of the organization and relevant decision-makers who can resource it.

Initiating a Responsible AI Practice

Before initiating a responsible AI practice, someone needs to recognize that such a practice is needed and push internally for staff and resources to build it. Sometimes this direction will come from the very top. For example, Microsoft President Brad Smith signalled responsible AI as a priority for the company in 2016 when he co-sponsored the formation of an internal AI ethics committee, [4] which worked to develop a set of public AI principles that the company introduced in 2018.[5] After laying these foundations, Microsoft created its first full-time role dedicated to AI ethics[6] and further institutionalized the work through the development of its internal Office of Responsible AI in 2019.[7]

Sometimes, energy may bubble up from an organization's research efforts. The team that research scientists Timnit Gebru and Margaret Mitchell developed and led at Google. was conceptualizing key concepts such as "model cards" in their research outputs long before the company had the people or processes to adopt those practices.[8] Such teams can also play a significant role in issue spotting and surfacing pressing concerns to leadership. For instance,

---

[4] Horvitz, E. (2018). Aether Committee. https://erichorvitz.com/Aether_Committee_Microsoft.htm; [perma.cc/8DG7-3LZG]

[5] Smith, B., & Shum, H. (2018). The future computed: Artificial intelligence and its role in society. Microsoft. https://blogs.microsoft.com/blog/2018/01/17/future-computed-artificial-intelligence-role-society/ [perma.cc/GS57-A27H]

[6] Davenport, T. H. (2019). What does an AI ethicist do? MIT Sloan Management Review. https://sloanreview.mit.edu/article/what-does-an-ai-ethicist-do/ [perma.cc/NS8X-4XQT]

[7] Horvitz, E. (2018). Aether Committee. https://erichorvitz.com/Aether_Committee_Microsoft.htm [perma.cc/8DG7-3LZG]

[8] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P ., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2018). Model cards for model reporting. FAT* '19: Conference on Fairness, Accountability, and Transparency. https://arxiv.org/abs/1810.03993 [perma.cc/WGB6-8EQN]

Gebru and Mitchell's team was one of the first to raise serious ethical concerns about the trend towards large language models (LLMs) — principled action that shined a light on some of the many challenges responsible AI practitioners face and which led to Gebru and Mitchell's contentious exits from Google in 2021.[9] Nevertheless, their efforts had helped prompt and laid the groundwork for Google to begin building out its responsible AI review processes.

At other times, senior management may be motivated to prioritize responsible AI based on their own exposure to the issues. For example, Meta's former director of Applied Machine Learning group, Joaquin Quinioñero-Candela, pivoted to build a team focused on responsible AI, establishing and leading the company's Society and AI Lab (SAIL) in 2018, with the support of Meta's then-CTO. The founding and focus of that team, which would later evolve into Meta's first Responsible AI product team, was prompted in part by Candela's own personal "reeducation" studying emerging research on issues such as algorithmic bias.[10]

Organizations may understand the connection between investment in responsible practices and advancing consumer trust, and make investments to support responsible AI to create conditions more conducive to adoption and innovation. And in some cases, primarily at newer AI companies such as OpenAI and Anthropic, concerns about responsible use of AI and AI safety may be baked into the founding mission of the organization — though they are still ultimately vulnerable to competing incentives.[11]

However they come to prioritize it, companies will always need initial champions of responsible AI development to help focus the attention of senior leadership, galvanize action, and integrate work into existing organizational structures and processes. Once those initial champions have succeeded in promoting investment in responsible AI, they're going to need to find a diverse and resilient team to implement their vision, whether it be assembled from existing staff, bolstered by outside hires, supported by external experts, or all of the above.

[9] Hao, K. (2020). We read the paper that forced Timnit Gebru out of Google. Here's what it says. MIT Technology Review. https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/ [https://perma.cc/F2S5-GRV6]; Metz, C. (2021). A second Google A.I. researcher says the company fired her. The New York Times. https://www.nytimes.com/2021/02/19/technology/google-ethical-artificial-intelligence-team.html [https://perma.cc/KQ58-5FKU].

[10] Hao, K. (2021). How Facebook got addicted to spreading misinformation. MIT Technology Review. https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/ [perma.cc/KS32-YCGV]

[11] OpenAI. (2025). https://openai.com/our-structure/ ("We founded the OpenAI Nonprofit in late 2015 with the goal of building safe and beneficial artificial general intelligence for the benefit of humanity."); Anthropic. Making AI systems you can rely on. https://www.anthropic.com/company [https://perma.cc/U9V5-LW4X] ("Anthropic is an AI safety and research company. We build reliable, interpretable, and steerable AI systems.")

Multidisciplinary Talent

Building a successful responsible AI practice depends on having practitioners with domain expertise – in relevant specializations, such as technical or qualitative research, product, and policy, or in a given field where AI might be deployed, such as healthcare or finance — who can work effectively with both technical and non-technical teams. Given the cross-functional and interdisciplinary nature of responsible AI, there are several design choices that will influence the long-term efficacy of these efforts.

- **Focus on encouraging interdisciplinary work rather than finding 'unicorn' employees.** One challenge to finding responsible AI talent is that practitioners are often expected to possess a rare combination of skills: technical expertise, legal fluency, communication capacity, management experience, and an ability to work in highly ambiguous environments. Yet effective work can happen when led by individuals and teams with a variety of strengths who are capable of collaborating effectively.

  "Basically, we want people with PhDs in computer science and law degrees and super great communication and management skills and creativity. Do these unicorns even exist?"

  Domain expertise in specific risk areas can ground conceptual notions of AI impacts in reality, ensuring that risks are properly calibrated and that focus remains on the issues that matter most as AI systems are developed in general or customized for particular domains.[12] As AI becomes more ubiquitous and is deployed across a variety of contexts, from healthcare to finance to agriculture, practitioners with deep contextual knowledge will need to work alongside professionals with deeper AI expertise. Whether this expertise is added through permanent hires, consulting contracts, or stakeholder engagement, it is critical to detecting and crafting meaningful mitigations for risks that are specific to those domains.

- **Ensure processes for acquiring talent do not structurally exclude necessary backgrounds.** Historically, many organizations have privileged technical expertise[13] in hiring for responsible AI roles. While technical skills can be important for certain roles, there are no responsible AI problems that do not involve people, and thinking about problems only technically — and prioritizing the involvement of technical staff — overly narrows the scope of work. Structural decisions in how to staff responsible AI efforts,

[12] For example, one study concluded that technical practitioners using "fairness toolkits" would benefit from deliberations with experts in law or gender studies. Deng, W. H., Nagireddy, M., Lee, M. S. A., Singh, J., Wu, Z. S., Holstein, K., & Zhu, H. (2022). Exploring how machine learning practitioners (try to) use fairness toolkits. ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). ACM. https://arxiv.org/abs/2205.06922 [perma.cc/J54F-URWV]

[13] Narrow definitions of 'technical' have been used in conversations on responsible AI when practitioners from fields other than computer science are quite technical experts on other subjects. For example, many social scientists have stronger understandings of advanced statistics than ML engineers.

including situating those efforts within technical parts of the organization, may limit the talent pool. For example, research scientist roles for responsible AI may fall under an organization's standard researcher job descriptions and require coding tests; in such a case, researchers from the social sciences and other disciplines, who have relevant research experience for responsible AI, may be disadvantaged or excluded.

Leaders instituting responsible AI efforts should be mindful of how hiring and staffing processes within their organization may undesirably limit the kind of people that can be tapped to support and drive this work. Leaders should also be aware that existing staff may look for practitioners who are like them — either in their expertise, educational background or employment history — and develop processes to manage potential blind spots when staffing responsible AI projects.

"It's really hard to hire non-CS [non-computer science] people when you have only interviewed CS people your entire life. How do you make sure that their expertise is evaluated fairly?"

- **Consider talent that may not have a direct background in responsible AI governance, policy, or research.** Individuals without explicit responsible AI work experience may still have ample relevant insight from exposure to user research, compliance, or policy.

  Companies can begin by empowering existing teams and functions to add responsible AI efforts to their portfolio. Organizations with fewer resources who may not be in a position to build dedicated responsible AI teams might consider designating a portion of their existing workforce's time to responsible AI efforts. For example, lawyers, UX researchers or compliance officers should be professionally rewarded for allocating a portion of their time to responsible AI efforts. While having dedicated staff is often more effective, the approach proposed here can help smaller organizations build their maturity without entirely redirecting staff from other priorities.

  "There are a lot of people in the organization who would love to help in [RAI] efforts, but are not sure where to start or where to go."

  Crucially, having the right talent is not enough to conduct responsible AI work or to demonstrate the commitment to doing so. Consistent with existing research on the subject, many practitioners we interviewed shared their experiences with unstable organizational structures, burnout, and frequent reorganization or dissolution of teams, all of which not only made the work of responsible AI harder, but also reflected poorly on the organization when such dynamics became public.[14] [15] [16] Organizations need to create a stable, healthy working environment for responsible AI practitioners, starting with effective working structures.

Effective Team Structure

Responsible AI efforts may be taken on by centralized teams, personnel distributed across teams, or some combination of the two. Centralized teams tend to have a broad remit and interact with a number of functions of the organization, from compliance functions to product or business teams to legal organizations. Distributed approaches involve embedding people across distinct parts of the organization, such as in specific product or business areas, where they may have a narrower scope but more integration within specific units. Each design choice has its strengths and weaknesses.

- **Consider whether your organization is better-suited for a distributed or centralized structure.** Distributing responsible AI efforts across product teams or business units can be effective in tackling specific concerns, especially those requiring greater contextual expertise, and the agility and flexibility of a distributed structure can enable organizations to more quickly address unique issues as they arise. However, with distributed teams, organizations may struggle to scale responsible AI efforts that would benefit from the development of common technical tools or the implementation of standardized and consistent approaches to decision-making. Distributed staff must contend with inter-team variations in goals, measurement methods, documentation practices, and approaches to responsibility, among other issues, which can lead to significant time spent coordinating and reconciling disparate efforts.

  A centralized responsible AI team may be better positioned to provide a unified infrastructure that can support sustained and consistent efforts across an entire organization. Such a team's remit could include leading the development of core responsible AI principles that apply across the organization to guide decisions, developing a coherent strategic approach to implementing responsible AI work and governance, establishing shared metrics, defining procurement guidance, and managing a clear distribution of labor across practitioners. Moreover, centralized teams may be able to define a mandate that is clearly supported by executives and decision-makers

[14] Heikkilä, M. (2022). Responsible AI has a burnout problem. MIT Technology Review. https://www.technologyreview.com/2022/10/28/1062332/responsible-ai-has-a-burnout-problem [perma.cc/XP8T-XXL5]

[15] Field, H., & Vanian, J. (2023). Tech layoffs ravage the teams that fight online misinformation and hate speech. CNBC. https://www.cnbc.com/2023/05/26/tech-companies-are-laying-off-their-ethics-and-safety-teams-.html [perma.cc/3BW2-DXN5]

[16] Schiffer, Z. & Newton, C. (2023) Microsoft lays off team that taught employees how to make AI tools responsibly. The Verge. https://www.theverge.com/2023/3/13/23638823/microsoft-ethics-society-team-responsible-ai-layoffs [perma.cc/8CPW-U5MP]

and will likely experience fewer siloes and more regular communication than distributed efforts.

However, centralized teams tend to have less specialized expertise and visibility into particular product or business areas, which may pose challenges as they work to meet the specific needs of other teams with whom they are partnering. This challenge can be exacerbated if product teams view centralized RAI teams as a costly barrier to launch or as interfering with their own decision-making. Additionally, in cases where responsible AI efforts are centralized, other teams or business units may lack the incentive to take on responsible AI work that may seem to slow down their progress, fall outside of their area of responsibility, or be in tension with their goals.

"Distributed responsible AI efforts are often time-bounded by project timelines and do not scale well, while centralized responsible AI may lead to bureaucracies that create barriers for addressing product-specific challenges."

Organizations may also consider a hybrid approach, such as a hub-and-spoke model, in which a centralized team actively collaborates with assigned representatives across various departments. For example, in some cases, experts and engineers in a centralized responsible AI function can establish foundational principles, priorities, and technical infrastructure, such as measurement tools and common mitigation techniques, while employees distributed across the organization, such as in particular product lines, incorporate those tools and principles within their domains.

While many technology companies made early investment in centralized efforts, institutions are increasingly moving toward decentralization. Having embraced a company-wide initiative in the early days of its responsible AI practice, Microsoft has since diffused some of that talent.[17] Meta has also moved toward a more decentralized approach, with responsibility distributed across specific teams.[18] On the other hand, some companies have shifted to integrate responsible AI efforts into centralized processes, presumably to reduce redundancy of a variety of related efforts.

Moving away from centralized efforts can erode a sense of centralized accountability — in the words of one practitioner, "if everyone is responsible for responsible AI, then no one is responsible." However, the organizational and operational advantages to distributed work can be significant. Ultimately, the most effective approach for an organization will depend on its size, maturity, existing leadership structure, and cultural

---

[17] Crampton, N. (2023). Reflecting on our responsible AI program: Three critical elements for progress. Microsoft On the Issues.
https://blogs.microsoft.com/on-the-issues/2023/05/01/responsible-ai-standards-principles-governance-progress/ [perma.cc/GUN5-ZW92]

[18] Reuters. (2023). Meta moves members of its Responsible AI team to other groups. Reuters.
https://www.reuters.com/technology/meta-breaks-up-its-responsible-ai-team-information-2023-11-18/ [perma.cc/UGE6-7NBG]

conditions, such as whether the institution sets overall goals in a top-down manner or empowers teams to set their own agendas.

- **Understand the advantages and disadvantages of separating responsible AI efforts from revenue-generating units in the organization.** Responsible AI efforts sometimes sit within cost centers in the organization; these are units that are not expected to generate revenue, and often include legal, compliance, and other risk management functions. When situated within a cost center, responsible AI practitioners might have less pressure to generate revenue, but their authority is likely more constrained, and they may also find it more difficult to obtain the resources they need to be creative, take risks, and demonstrate value to the organization.

  Proximity to a company's core offerings may influence the efficacy of responsible AI efforts. On the one hand, when situated within revenue-generating parts of the company, such as a product team or other business units, responsible AI teams are more directly exposed to meaningful business challenges and opportunities, which may make it easier to have a direct impact on final outcomes of product decisions. They may also be taken more seriously because they are perceived to be closer to functions that drive value for the company.

  On the other hand, situating those driving responsible AI objectives within a business unit can make them less independent, and leave their work vulnerable to pressure from leaders eager to ship products quickly or maximize business metrics. Organizations should carefully consider which functions are best suited for central efforts and when and how personnel embedded in other relevant teams across the company ought to take on responsible AI work.

Effective Collaboration and Engagements

To ensure responsible AI practices are effective and meet the expectations of outside stakeholders, practitioners and teams should prioritize regular engagement with both internal and external voices with interest and expertise in ensuring that AI risks are appropriately managed. When collaborations with outside experts are formalized, they can offer opportunities to align responsible AI efforts with broader community consensus, translate responsible AI efforts into concrete products and technologies, and imbue ongoing work with expert insights.

- **Set up structures for internal engagement that set clear expectations, and ideally encourage collaboration.** Because responsible AI teams and practitioners are unlikely to "own" products or be experts in all domains in which their work is relevant, relationships with peers across their institution are instrumental to motivating, guiding, and embedding responsible practices into product development and deployment. Defining structures like office hours, regular check-ins with partner teams, informal reviews, or formal checkpoints can improve clarity on how teams can engage with responsible AI efforts. These structures may not always be maximally efficient, but they are nevertheless valuable in helping responsible AI practitioners understand their colleagues' challenges and priorities better, and, in turn, advancing responsible practices more effectively.

These structures do not emerge by default; they must be defined, established, promoted, and staffed, with clarity about what happens before, during, and after an engagement between responsible AI structures and other teams. For instance, a central responsible AI team can create structured intake forms to collect relevant information about an AI use case to facilitate a consultation, and commit to providing responsive recommendations on a defined timeline. A team member may be designated to capture details of the case and the resulting recommendation in a central repository in order to inform future consultations on similar topics, and a defined escalation path can be established so teams know which process to follow in the event of a disagreement over recommended practices.

The choice of internal engagement mechanisms will depend heavily on which structure an organization has chosen for responsible AI efforts (e.g., centralized, distributed, or hybrid). An organization with a centralized team and clear compliance obligations, for instance, may have a formal process to review teams' work or independently replicate tests of a product or model, reverting to a team with a clear go/no-go decision or a set of minimum requirements to proceed. Defining and creating organizational infrastructure to support the preferred mechanisms of engagement will help ensure these processes are effective. Personnel like program managers can be instrumental in facilitating these processes and coordinating relevant internal stakeholders.

"It can be intimidating to come to a 'council' or even a team meeting with strangers who are presumably there to criticize your work, so sharing template expectations or a 'what to expect' up front and framing those meetings positively is really important."

Many practitioners agree that the manner, tone, and style of engagement is instrumental to building trust; structures where certain personnel or teams are perceived to be operating without sufficient authority or expressing their recommendations forcefully without clear leadership backing can be seen as illegitimate.

On the other hand, where expectations have been clearly set and responsible AI practitioners have a mandate to hold other teams across their organization accountable against defined requirements, a more formal relationship and related fora for adjudicating decisions may be appropriate.

- **Establish thoughtful and ongoing structures to engage with external experts and stakeholders — and be cautious of engagements that come across as performative.** Researchers, advocates, and policymakers generally agree that engagement with experts and impacted communities is paramount for responsible AI and risk management efforts to be meaningful. But conducting these conversations in an ad-hoc and irregular manner can lead everyone to be dissatisfied by the experience — practitioners may not get the feedback they need and those consulted may feel taken advantage of or tokenized. External experts often juggle multiple requests to engage with companies and organizations, and in some cases may be working with multiple teams within the same institution, so establishing a relationship owner who can build rapport with relevant experts and communities and establish preferred modes of engagement can be useful to build channels for sustained and constructive feedback.

To navigate such challenges, some organizations opt to set up advisory councils that meet on a regular basis and develop familiarity with an organization's main products or issue areas (which requires resources both to coordinate logistics and terms of engagement and to compensate experts appropriately for their involvement). Others opt for concentrated, intensive engagements like co-design workshops and citizen juries, which can facilitate deeper collaboration and context-sharing. Some seek engagement with industry-specific consortia and bodies. Still others build dedicated tools for researchers to access data and information about systems in order to facilitate longer-term research into and evaluation of AI systems.

Whichever choice is made, organizations should be careful to avoid versions of such engagements that err on being (or appearing to be) overly performative, for example, by seeking high-level and thematic feedback but giving people insufficient detail about what the company or organization is building, making no commitments to incorporate the feedback participants provide, failing to compensate people for their contributions, or not providing accommodations so that such events can accessible to the communities they intend to reach. Assigning responsibility for engagements and stakeholder relationships and setting up infrastructure and funding to facilitate them can help augment the quality of feedback that external experts are in a position to provide and avoid known pitfalls.

All of the practitioners we spoke to urged investment in meaningful dialogue with key external experts and stakeholders. That said, a number of them raised skepticism about recent "democratic AI" proposals that centered industry players and products. Several voiced the concern that these broader efforts are misguided in their focus on building broad new deliberative processes rather than supporting interventions by existing representative democratic processes. Others went further to argue that this was in fact the point: that such efforts are intended to distract from and hedge against regulation. But several defended these efforts as a necessary innovation in engaging broader communities and concerns, believing that more conventional approaches were inadequate for keeping pace with the rate of technological change.

Priorities: Thoughtfully Triage Work

Irrespective of which sector a responsible AI team is working in, it will need to define a portfolio of work, set up mechanisms for establishing and managing priorities, and develop a strategy for execution, fundraising, and people management.

Portfolio of Work

An important step in institutionalizing responsible practices is demarcating what this work entails, and how it is distinct from (and/or integrated into) compliance, legal, and other related efforts. Early responsible AI efforts in the tech industry were typically housed within research teams and were distinct from, though in some cases informed by, more formal compliance and legal processes. Emerging regulatory requirements and changing technological realities, including the surge in popularity of generative AI, have impacted both the relationship between responsible AI efforts and compliance teams as well as the kinds of expertise and requirements the portfolio depends on.

- **Anchor responsible AI efforts in regulatory requirements — but not to the exclusion of other concerns.** Grounding responsible AI in existing regulation and norms, such as international human rights, can help these efforts meet external expectations, since these frameworks are typically the result of in-depth negotiation among a variety of stakeholders. Regulations can be helpful in motivating and institutionalizing sustained responsible AI work since regulatory requirements can make it easier to convince various stakeholders across a company to integrate responsible AI practices.[19] Even if it does not end up being the sole or primary driver of responsible AI work, regulation can be a powerful tool to shape the focus of responsible AI efforts. (Other policy-informed, consensus-driven resources like the NIST AI Risk Management Framework and related guides, such as the Generative AI Profile, as well as resources like the UN Guiding Principles on Business and Human Rights, can also be a helpful starting point for organizations to begin scoping their responsible AI efforts.)

  "It's a lot easier to get buy-in for the compliance problem than for the ethics problem."

  However, organizations should be aware that incorporating or subsuming broader work on potential user and societal harms under compliance efforts may dilute or appear to dilute these ethically-oriented efforts. Research suggests compliance requirements do

[19] Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2021). Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. Proceedings of the ACM on Human-Computer Interaction, 5. https://arxiv.org/abs/2006.12358 [perma.cc/N2HD-EKKR]

not necessarily lead to ethical behavior, and ethical progress may not mean increased compliance.[20]

Practitioners echoed these concerns, reflecting that a compliance-focused mindset risks a focus on complying with external rules and standards without engaging underlying ethical concerns.[21] For example, it might lead a company to narrowly define and measure bias when a broader lens involving deeper research within a particular context may be necessary. If an organization opts to prioritize only work that is required by law or regulation, ethics-focused research or projects might be underfunded, underappreciated, or absent.

"Some lawyers think their job is to help the company meet a minimum; their incentives might be different from those of responsible AI practitioners."

- **Build channels to gather practitioner expertise on what additional emerging issues merit attention.** Often, practitioners closest to planned products and users can surface high-priority issues and identify effective solutions. When responsible AI practitioners are empowered to spot and highlight issues that are clearly presented by technology and products their organization is pursuing, they may be able to tackle those issues earlier and prevent more serious impacts for users, communities, or their organization.

  For instance, practitioners noted that some organizations had started investing in exploratory work into demographic performance gaps and other issues in systems like voice assistants, even before external pressure developed, because the teams building those products observed early how they could fail for certain consumers. On the other hand, if these embedded efforts remain ad-hoc and are not communicated to central responsible AI efforts, risk management processes, or relevant decision makers, it may be challenging to mobilize colleagues to learn from and embed best practices across the organization. And without strategic support, such efforts can risk being — and, indeed, often have been — deprioritized in favor of other 'fires'.

- **Be open to refining the scope of work.** In the face of shifting expectations and requirements, organizations may need to consider how much responsible AI efforts should be grounded in the (i) business'understanding of the risks in what they are building or deploying, (ii) user-research-informed considerations of impact on users and the public, (iii) focus areas recommended by external stakeholders, or some combination of the three. In practice, this scoping exercise may depend on why responsible AI is being prioritized within an organization at any given moment. Common incentives that

[20] Michaelson, C. (2006). Compliance and the Illusion of Ethical Progress. Journal of Business Ethics. https://link.springer.com/article/10.1007/s10551-005-5589-8 [perma.cc/3LK8-4XGS]
[21] Casiraghi, S. (2023). Anything new under the sun? Insights from a history of institutionalized AI ethics. Ethics and Information Technology. https://link.springer.com/content/pdf/10.1007/s10676-023-09702-0.pdf [perma.cc/W33V-P56Z]

drive technology companies include harm avoidance, motivation to establish or maintain legitimacy to operate, public relations risks, and the desire to avoid regulatory or legal risk.[22]

Priorities

Initial scoping exercises are likely to surface a long list of potential focus areas and practices — more than will be possible to tackle all at once, even for well-resourced teams. Prioritizing among them is an ongoing challenge: even after a robust responsible AI practice is established, practitioners need to continue balancing evolving and competing goals, pressures, and deadlines. Moreover, responsible AI practices, like trust and safety efforts, are typically established well after a company's product has been developed, so teams often face a backlog of existing issues and technical debt.

"What external folks might not realize is that we are presented with hundreds of risks and have to ask: which 10 should we be doing right now? What rises to the top? Should it be the legal angle that takes priority or the most recent PR angle?"

- Prioritize severe and urgent issues, but don't ignore longer-term investments. Especially for organizations with fewer resources, it may make sense to tackle one or a few issues at a time instead of trying to accomplish everything at once. This requires triaging and prioritizing risks based on severity and urgency while also allocating resources for longer-term risks that otherwise risk being consistently deprioritized.

  Responsible AI leaders will need to weigh considerations like team capacity, especially given the studies showing that practitioners commonly experience burnout.[23] Leaders also need to navigate competing priorities that partners in business and product teams may be facing, which may constrain the latter's ability to contribute to collaborate on detection and mitigation efforts.

  "Among RAI folks, RAI is the sun and the moon and the stars. You move these ideas to product teams, and they tell you RAI is one of 17 things they're doing."

  When making prioritization decisions, key factors to consider will often include:
    - Explicit regulatory requirements
    - Product-related concerns, including frequency and severity of predicted harms and scale and sensitivity of use-cases
    - Emerging regulations and regulatory attention on key issues

[22] Bogen, M. (2016). Wired for geopolitics: Incentives shaping technology companies' international policy decisions (Master's thesis, The Fletcher School, Tufts University).
[23] Heikkilä, M. (2022). Responsible AI has a burnout problem. MIT Technology Review. https://www.technologyreview.com/2022/10/28/1062332/responsible-ai-has-a-burnout-problem/ [perma.cc/7HBQ-VRQJ]

- Consumer-reported harms
    - Considerations of societal impact, (e.g., by considering human rights frameworks)
    - External and internal expert concerns, including from academia and advocacy groups

- **Build processes to ensure 'important but non-urgent' work is not forgotten.** Crucial issues that are less urgent or may present risks on a longer time horizon can be harder to manage because they may lack obvious mitigations or require ongoing attention. Moreover, only focusing on what the organization knows how to do can excessively narrow the scope of responsible AI efforts. This, in turn, can lead to missed opportunities for addressing larger systemic issues.

  For example, several AI companies have prioritized testing for explicitly discriminatory decisions, but do not seem to have proactively explored the risks that more subtle biases might pose or how to address them. For instance, Anthropic conducted research on whether Claude 2.0 exhibited discriminatory behavior on decisions related to insurance, credit, welfare, criminal justice, and other consequential contexts when protected characteristics are explicitly mentioned,[24] but third party research has demonstrated that stereotypes can be more insidious, such as associating African American Vernacular English with less prestigious jobs and harsher courtroom convictions.[25]

  One approach practitioners can consider involves starting with a specific problem, conducting a root-cause analysis, and identifying both easy fixes that can be adopted more quickly as well as deeper systemic issues. This exploration can lead to unexpected solutions as well as identify "low-hanging fruit" efforts that can be applied broadly to address more common issues, freeing up time to dig into more complex challenges for which research may still be needed to develop measurement and mitigation approaches. Companies can also prioritize initiatives that enable external researchers to study AI use cases to foresee likely harms or spot underappreciated ones.

- **Don't ignore civil society and researcher perspectives on which risks and concerns to prioritize.** While it is appropriate for companies to track and be sensitive to regulatory signals, risks and concerns raised by civil society and researchers, as well as other external experts, still require attention. Unfortunately, companies often deprioritize such voices and issues even before they have been fully addressed. For example, while Meta vocally prioritized AI fairness and civil rights for several years, the company seems to have diverted efforts to tackle bias and discrimination, even though stakeholders

[24] Anthropic. (2023). Evaluating and mitigating discrimination in language model decisions. https://www.anthropic.com/research/evaluating-and-mitigating-discrimination-in-language-model-decisions [perma.cc/X39M-5GKH]
[25] O'Grady, C. (2024). AI makes racist decisions based on dialect. Science. https://www.science.org/content/article/ai-makes-racist-decisions-based-dialect [perma.cc/5GTQ-RXMJ]

continue to point out the effect those harms were having on the company's users.[26] The company's public documentation for their Llama 3 model, for instance, does not address bias and fairness.

Strategy

Responsible AI practitioners need to develop a strategy that enables the work to be executed effectively, including securing support from leadership, acquiring necessary resources, and navigating organizational dynamics.

- **Find and sustain support from senior leadership.** While it is generally agreed that leadership buy-in for responsible AI efforts is crucial, what effective leadership engagement looks like can be less clear.

  First, practitioners should leverage senior leaders and colleagues they trust to secure resources and approve prioritization of responsible AI work. In the technology sector, vice presidents (VPs), senior leaders who can wield both hard and soft power, can be particularly important figures to drive and sustain this work. VPs have status within an organization as a result of their seniority and perceived talent, and are often responsible for a variety of teams, with the ability to define priorities and allocate substantial resources within their remit. In particular, leaders of product or business teams are key allies because they tend to hold full responsibility over their products and are empowered to make significant decisions within their domains — including whether to prioritize responsible AI efforts, and how to adjust timelines and deliverables to accommodate this work.

  "Many people across the AI lifecycle want to do good things, but if they don't have the buy-in, none of it sticks."

  Second, mid-level leaders such as engineering or business leads should be informed and empowered to support responsible AI efforts. If these individuals are unaware of or disagree with the direction of responsible AI efforts, they may pose considerable friction to internal efforts even when executive buy-in exists. While these employees typically report to VPs or similarly-situated business leaders, they often have ample work and management responsibilities without the power and prestige of being a senior executive, and so may opt not to prioritize responsible AI efforts in practice.

[26] Meta. Llama 3.1 model information. https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md [https://perma.cc/6GLH-Q47E]; Llama Team, AI @ Meta. (2023). The Llama 3 herd of models. arXiv. https://arxiv.org/pdf/2407.21783 [https://perma.cc/9VZH-3AQ4]

- **Demonstrate sensitivity to organizational dynamics and culture, and find ways to make them work for responsible AI efforts.** To be effective, responsible AI teams will need to adeptly navigate organizational dynamics, including competition between teams, infighting between senior leaders, and potentially challenging cultural circumstances, such as an expectation that employees be available at all hours.

  For example, practitioners may regularly need to justify their budget and make the case for responsible AI to continue being among the organization's priorities. In an organization that grounds its work in a set of values or code of conduct, responsible AI practitioners might do well to root the responsible AI principles and practices in those structures.[27] RAI practitioners in an organization that is customer-driven will want to focus on finding user stories that demonstrate the harms they aim to address.[28] And in a product-focused organization they may require strong allies that understand product roadmaps, priorities, and infrastructure.

Processes: Establish Structures for Governance

To avoid ad-hoc decision-making and reinventing the wheel with each new substantive or procedural governance decision, organizations need to set up mechanisms, processes, and guidelines for problem-solving and issue resolution. While it may seem that a robust culture of risk management will slow work down, when executed correctly it can enable the organization to move faster in developing and deploying technologies or use cases for which there are already clear guidelines. These mechanisms include developing sector-specific risk management practices; implementing guidance for personnel within the organization who will need to make decisions on a regular basis; and establishing checks and balances to ensure incentives are aligned and that relevant stakeholders are engaged on important decisions.

Risk Management Approach

Responsible AI practices are rooted in an understanding of which risks need to be eliminated, managed, monitored, or tolerated. Industries such as aviation, finance, and healthcare have developed the context-specific vocabulary and framework for risk management over decades of practice, research, and debate. AI risk management is still emerging, and little consensus exists even on foundational questions. However, the following considerations will be useful across contexts.

---

[27] Microsoft. Corporate Responsibility. https://www.microsoft.com/en-us/corporate-responsibility [https://perma.cc/MSB2-USR8]
[28] Google. Commitments. https://about.google/intl/ALL_us/commitments/ [https://perma.cc/Y3WB-Z9CM]

- **Standardize the work of risk management to the extent possible to ensure decision-making is consistent and repeatable rather than ad-hoc.** In heavily regulated industries, like banking and healthcare, sensitive and high-risk uses of AI are subject to standardized risk management approaches. In domains where such risk management approaches may not yet be legally required, governance decisions on the use, deployment, and monitoring of AI tend to be ad-hoc and inconsistent. Emerging regulations including the EU AI Act aim to close this gap,[29] but AI risk management remains a nascent field. Moreover, in the absence of robust AI regulation, many companies are tolerating extremely high risks for the sake of competitive advantage.

  "Risk tolerance in AI is so high no matter how you calculate it; we're talking about possibly trillions of dollars at stake for moving slowly."

  A useful place to start, especially for an organization at the early maturity stage of developing its approach to AI risk management, is mapping relevant risks that AI may pose. This involves looking at the various development lifecycles in the organization — including data, models, and product, as well as procurements from external vendors — identifying salient risks posed by each stage, determining the prevalence and magnitude of those risks, and building and applying risk mitigation and management strategies.

  Piloting particular risk mitigations within one product or team can be a useful way to assess whether they're effective and demonstrate the value of scaling to the whole organization. Ad-hoc engagements can allow practitioners to develop meaningful relationships with product and engineering teams that are personalized to these teams' needs and idiosyncrasies, which can support the development of more sustained risk management practices

  For more mature organizations, emerging AI standards such as ISO/IEC 42001 and ISO/IEC 23894 can be a useful starting point.[30]

- **Determine which risks need to be minimized, mitigated, or tolerated, and manage them appropriately.[31]** A clear sense of an organization's risk tolerance can help to streamline decisions related to responsible AI priorities and requisite resourcing; the sooner organizations can define their appetite for different risks, the more straightforward implementation will be. When teams understand clearly what risks must be addressed prior to launch, they can build that work into their timelines, budgets, and work plans from the start rather than realizing at the last minute that certain risks have

[29] For a thoughtful literature review on current approaches to risk management in AI, see Jedličková, A. (2024). Ethical considerations in risk management of autonomous and intelligent systems. Ethics & Bioethics. Sciendo. https://sciendo.com/article/10.2478/ebce-2024-0007. [perma.cc/WGC7-3R5R]
[30] Responsible Artificial Intelligence Institute. (2024). Putting AI standards into action. https://www.responsible.ai/putting-ai-standards-into-action/#:~:text=Focus%20first%20on%20well%2Dknown,are%20also%20certifiable%20and%20auditable. [perma.cc/M2WG-4WN5]
[31] Helpful resources include IEEE Standard for Software Life Cycle Processes and ISO 31000

not been sufficiently reduced. Managing risks will inevitably involve trade-offs, so creating the conditions to begin the process sooner will empower teams to spend their time and resources most wisely.

Practitioners must recognize, though, that risk assessment is not a neutral or objective process, but one that reflects the values and perspectives of a specific community.[32] Organizations must make explicit choices about which risks to measure and prioritize, and develop frameworks for how to balance given risks with other goals as well as how to adjudicate which risk takes priority when two conflict. If companies overly defer to legal and compliance obligations to guide this assessment, negative externalities such as impacts on communities or the environment may be sidelined. To reduce the likelihood of this imbalance, signals like customer pressure to address negative experiences, brand impact and reputational risk of potentially irresponsible choices, and normative commitments a company has made should also be factored into risk assessment exercises.

- **Establish risk tolerances grounded in data and evidence, as opposed to speculation.** Many practitioners underscored the importance of taking an evidence-based approach to risk in order to calibrate which warrant attention, but acknowledged how difficult this is to do; there is no universally-accepted instruction manual for setting appropriate risk thresholds in AI.

  An overly restrictive sense of risk can be nearly as detrimental as an overly permissive one. Practitioners in government and public services expressed the opposite problem to those working in the technology sector: risk tolerance tends to be so low that it can be difficult to innovate on key areas that would contribute to overall public good, including in developing creative approaches to addressing risks. If practitioners are overly concerned with narrow compliance requirements, this can make it difficult to deliver important services to the public. Moreover, perceptions that responsible AI is a source of friction rather than an effort to meaningfully help people can dissuade people from engaging with the topic. Where evidence of certain risks may be sparse, practitioners can start by investing in practices to help gather the information necessary to inform sound risk management decisions.

Checks & Balances

Ideally, responsible AI work would be self-motivated, sufficiently resourced, and its practitioners rewarded within an organization's standard incentive structures. But in reality, the strongest motivators for this work tend to be avoiding reputational risk, enhancing customer trust, demonstrating compliance with existing regulations and trying to shape new ones. If making the case to organizational leadership for sufficient investment in responsible AI efforts requires

[32] Krebs, J.R. (2011). Risk, uncertainty and regulation https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2011.0174 [perma.cc/FL4F-4E3X]

clearly situating them in relation to the organization's bottom line, it is particularly important to establish safeguards to prevent the work from ending up subject to a set of perverse incentives.

- **Work to establish a structural balance of power to ensure that responsible AI work proceeds effectively and with integrity.** Organizations should be structured in a way that aligns incentives and goals.
    - Insulate decision-makers on product and risk from other parts of the organization, such as public affairs or communications, to ensure they have sufficient autonomy for their work.
    - Create reporting lines to relevant senior leadership or the board that balance power and ensure different parts of the organization are appropriately incentivized.
    - Consider lessons from the governance models of regulated organizations like banks and airlines (e.g., establishing the 'Three Lines of Defense'[33]).

  The Three Lines of Defense model involves the separating employees into the "first line", which is responsible for business outcomes and profits, including products and services for customers, and their associated risks; the "second line", which evaluates, challenges, and supports the risk management efforts conducted by the first line; and the "third line", which includes internal audit functions that can more independently assess residual risks and ensure the organization is meeting its obligations.

  While this model is common across many established enterprises, it remains rarer in the context of responsible AI.[34] However, companies like Microsoft and Alphabet have some risk management practices that are consistent or overlap with the model, such as the internal audit function.[35] In many organizations, fully implementing the Three Lines of Defense model may not always be feasible. Yet the spirit of the model can be informative in navigating organizational dynamics that may undermine responsible AI work, and support the building of structures to manage them.

The most effective structure may vary across organizations; for instance, having both a Chief Operating Officer (COO) and a Chief Technology Officer (CTO) responsible for reporting on AI governance to the CEO and the Board creates a double layer of oversight — more than one senior leader has responsibility over the practice. Such a structure can help ensure that the departure of or shifts in one's leader's priorities are less disruptive to the continuity of responsible AI efforts, and may help prevent ethics washing or a narrow agenda informed by just one leader's views. That said, a number of practitioners highlighted that having more than

---

[33] Schuett, J. (2023). Three lines of defense against risks from AI. AI & Society. https://link.springer.com/content/pdf/10.1007/s00146-023-01811-0.pdf [https://perma.cc/3PBN-CD8L]; Robinson, B. & Ginns, J. (2024). Transforming risk governance at frontier AI companies. The Centre for Long Term Resilience. https://www.longtermresilience.org/wp-content/uploads/2024/07/Transforming-risk-governance-at-frontier-AI-companies-CLTR-1.pdf [https://perma.cc/6HT7-CYSS]

[34] Schuett, J. (2023). Three lines of defense against risks from AI. AI & Society. https://link.springer.com/content/pdf/10.1007/s00146-023-01811-0.pdf [perma.cc/J47L-BU7R]

[35] *Ibid.*

one senior leader involved may create tension and competition between leaders, so organizations should consider broader dynamics and corporate culture before deciding which structure is most conducive to good work in this arena. What's critical is to develop a structure that enables the right checks and balances within the specific organizational context.

- **Create the right incentives for employees.** Tools, policies, and processes alone do not guarantee that responsible AI work will happen to the degree necessary. While it might seem that simply articulating the importance of governance efforts will lead to their adoption, technologists tend to face competing priorities that mean they are insufficiently incentivized to take these efforts as seriously as stakeholders hope. Moreover, technologists are too often incentivized to take short-cuts to avoid the time and resource costs of governance work.[36]

  To create the right incentives for practitioners to prioritize, organizations should consider:
  - Making the often invisible work of responsible AI visible by developing measurements and metrics for effective governance
  - Introducing sensitivity to risks as a formal evaluation metric in performance reviews, similar to those done in some regulated institutions
  - Formalizing governance requirements in product launch timelines, such as clarifying required check-points
  - Creating pathways for professional advancement of practitioners who take on responsibilities for AI risk management

- **Invest in both formal and informal mechanisms to advance and embed responsible AI.** Incorporating both formal levers, like legal and compliance obligations, and informal ones, like ethical commitments and workforce culture, is critical to the efficacy of a responsible AI practice, especially in instances where formal mechanisms struggle or fail.[37]

  Responsible practices, whether conducted under the auspices of formal compliance efforts or motivated by other choices within an organization, are enacted by people, rather than merely executed. Even in the face of clearly defined rules and procedures, culture and values determine whether these processes are diligently followed or aggressively avoided. Moreover, compliance efforts may be frequently negotiated and

---

[36] Winecoff, A. & Bogen, M. (2024). Improving governance outcomes through AI documentation: Bridging theory and practice. Center for Democracy & Technology https://cdt.org/insights/report-improving-governance-outcomes-through-ai-documentation-bridging-theory-and-practice/ [perma.cc/FN2H-GZSY]

[37] Papyshev, G. (2024). Governing AI through interaction: situated actions as an informal mechanism for AI regulation. AI and Ethics. https://link.springer.com/article/10.1007/s43681-024-00446-1 [perma.cc/5KV5-J7X4]

reinterpreted,[38] particularly as new technologies emerge and new risks are surfaced. Given this context, informal channels remain instrumental in helping practitioners take initiative to identify and address issues that may fall outside of clear requirements, follow the spirit, and not just the letter, of laws, and seek to minimize harms for customers and communities.

Ideally, both formal and informal approaches are integrated into hybrid implementation frameworks that contain both regulatory requirements and industry best practices. Without the formal elements, responsible AI and governance work is subject to ad-hoc changes, and may lack "teeth" or sufficient enforcement mechanisms to convey its importance. But without the informal elements, responsible AI implementation risks becoming a series of checklists.

Platforms: Invest in Responsibility Infrastructure

In developing an effective responsible AI practice, organizations should ensure they have the infrastructures, methods, and common vocabularies needed to operate effectively, and build on these foundations to improve robustness over time.

Foundational Infrastructure

For responsible AI efforts to move from ad-hoc research efforts to institutionalized practices, organizations typically need to develop procedural and technical infrastructure to facilitate at least the most common governance activities required of product or business teams. Common tasks that require supporting infrastructure include:

- **Model inventories.** In order to ensure AI systems are subject to relevant governance processes, a centralized inventory of models (whether deployed in-house or procured) and their related components and documentation lays a critical foundation for any risk management work. Standard artifacts like datasheets, model cards, system cards, and documentation of key decisions and risk controls can then be appended to a single record.
- **Measurement and evaluation tools.** Common infrastructure can ensure that measurements are implemented correctly and that results are displayed in standard formats or with appropriate statistical designations. Embedding common measurements in standard infrastructure can be particularly important for measurements that require

[38] Pérezts, M. & Picard, S. (2014). Compliance or comfort zone? The work of embedded ethics in performing regulation. Journal of Business Ethics. https://link.springer.com/article/10.1007/s10551-014-2154-3 [perma.cc/PRX5-ZG2H]

access to data that may be restricted to certain functions, such as sensitive demographic information, because the tooling can enforce access controls, required aggregation policies, or other privacy or safety guardrails.

- **Off-the-shelf prevention and mitigation methods.** Methods like de-identification, processes for re-sampling of training or evaluation data to ensure representativity, classifiers to detect certain types of data or behavior, and tools to facilitate model interpretability can be built into centralized infrastructure so different teams can manage common risks.[39]
- **Post-deployment monitoring tools.** Specific software tools for monitoring AI model performance and outcomes can help systemize the ability to spot issues once AI-powered products are deployed. These tools can help monitor evaluation metrics, enforce thresholds for when a model should be retrained, fine-tuned, revised, or completely shut down, and ensure other safety guardrails remain in place. Organizations procuring or modifying AI systems should establish a process to direct learnings back to vendors and catalyze upstream mitigations where necessary.

Several practitioners identified that embedding canonical expertise in existing infrastructure is important to help scale responsible AI work, since it allows non-expert teams to use available tools rather than relying on responsible AI practitioners to conduct required work on their behalf. Doing so also prevents circumstances where decisions about foundational practices are constantly revisited by staff who may lack the context, expertise, or authority to redefine those practices. One practitioner referred to the imperative of constraining “philosopher-king engineers” who consider themselves qualified to make their own decisions about complex responsible AI issues based solely on their technical and business expertise.

An open question for practitioners to consider is how much AI governance work their organization should attempt to automate. On the one hand, certain tasks related to responsible AI may be well-suited for automation — particularly if they are repetitive and don’t require significant substantive engagement, such as monitoring model drift or other changes. Identifying good candidate tasks for automation can help support organizational buy-in by reducing what can appear to be needlessly manual, expensive, and time-intensive processes. On the other hand, automating responsible AI work removes the opportunity for moral reflection and the development of particular skills, and may also lead to cases where poor decisions are made because the system in place lacked human judgment or common sense.

Shared Taxonomies

Responsible AI practitioners need a common language to communicate effectively, coordinate activities, and prioritize. To that end, developing shared taxonomies can be a useful exercise, even if it can feel more distant from the work of addressing known risks.

[39] Pair. Tools & platforms. https://pair.withgoogle.com/tools [perma.cc/9YAA-FEZM]

- **Define key concepts, like "fairness" and "transparency", and establish how to instantiate them into practice.** Because these definitions are foundational to other responsible AI policies, practices, and tools, progress on work like developing metrics, building mitigations and assessing performance can face significant friction without them. Just as in public debates about AI, such definitions are often hotly contested and difficult to align on internally[40], so teams should allocate enough time to reach a working consensus — and to capture the conclusions of that process. While there is an upfront cost to navigating these disagreements, failing to establish shared definitions can end up wasting time as practitioners talk past one another[41] or lead teams to assume they are working on similar challenges but later realizing they were working at odds with one another.

  "Although it constantly made me want to slam my head against my desk, I wish we had spent even more time debating the basic language we were going to use across the org to define our shared RAI goals. Every minute we didn't invest earlier in finding that precise language led to hours of argument later in the process."

- **Use common taxonomies to coordinate activities across teams and functions.** Risk and harm taxonomies, for instance, can help organizations clearly articulate the categories to be considered as part of AI governance practices, and help integrate AI-specific efforts with related risk management practices that consider similar risks and harms. Without defined taxonomies, disparate practitioners and teams often end up unintentionally duplicating or diverging in their efforts. Reconciling multiple frameworks can divert teams' attention for weeks or months, further delaying substantive work of addressing the most important risks and harms.

- **Use shared definitions and taxonomies to facilitate prioritization.** The existence of shared definitions and taxonomies does not answer the question of how to prioritize among different issues. But without these foundational resources, such prioritization conversations prove to be extraordinarily difficult because stakeholders and decision-makers lack a coherent understanding of what they are deciding on in the first place.

---

[40] Deng, W. H., Yildirim, N., Chang, M., Eslami, M., Holstein, K., & Madaio, M. (2023). Investigating practices and opportunities for cross-functional collaboration around AI fairness in industry practice. ACM Conference on Fairness, Accountability, and Transparency. ACM. https://arxiv.org/abs/2306.06542 [perma.cc/ADH4-CPVC]
[41] Lechterman, T. (2023). The concept of accountability in AI ethics and governance. In J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson, A. Korinek, M. M. Young, & B. Zhang (Eds.). The Oxford handbook of AI governance. Oxford University Press.

Implementation Guidance

Foundational definitions and taxonomies of common concepts are important starting points, but more bespoke guidance is often needed to help make decisions in the face of ambiguous fact patterns, issues in tension with one another, or competing priorities.

For instance, take an organization aiming to prevent biases in its AI systems. It may first need to determine which among all AI models or use cases require measurements to be conducted; organizations commonly prioritize such work according to which systems are likely to contribute to consequential decisions about people, or that may result in broad representational harms. Teams may wonder how much effort they are expected to invest in reducing bias, particularly if their early efforts appear to have complex effects, so defining thresholds, relative improvements, or specific actions that are expected to be taken can make these expectations clear from the start. Finally, the organization will need to decide how to proceed. It may compare several different versions of a model, finding that one version reduces bias for users with low socioeconomic status but appears to increase it for women, while the other results in the opposite effect. When deciding which to proceed with, it may need to consider whether to take a "do no harm" approach where they deploy a model that results in the least bias for any group, or whether to adopt a model that leads to the largest fairness improvements for the most groups — either way, having a schema by which to make these common decisions will make efforts more predictable and consistent.

More generally, responsible AI practitioners may be asked for guidance on:

- How to prioritize among a set of risks or harms;
- How to prioritize among a collection of product or application areas;
- How to determine what types of risks require the implementation of which mitigations or controls, and what to do if those risks or controls are not yet straightforward to implement;
- How to proceed if measurement indicates contradictory signals (e.g. reducing harms in one context appears to increase them in another);
- How to determine when a risk has been sufficiently mitigated in order to proceed with development or launch;
- How to document and demonstrate risk management efforts.

Despite best efforts, synthesizing responsible AI considerations into clear frameworks can risk oversimplification of concepts, leading guidance to be quite abstract and difficult for non-experts (or even experts) to parse. For example, emerging AI risk management frameworks like those from NIST provide some helpful procedural guidance on how to structure an overall AI risk management process, but offer little guidance on how to do that work or make critical decisions.[42]

Against this backdrop, more descriptive or narrative case studies that walk through particular products or fact patterns can help make these frameworks more concrete. Within an organization, early planning to maintain a repository of past decisions that practitioners can navigate to find cases similar to their own can help to centralize key learnings and internal “case law” that supports consistency and clarity as AI governance efforts become increasingly institutionalized.

These internal case studies can also prove useful as foundations for engagement with external stakeholders, both to gather early feedback on how to develop decision-making frameworks and how to navigate decisions in particular contexts, as well as to communicate about these efforts and their complexity to the public. Experts will expect to see, if not some degree of consistency across cases, at least a clear understanding and justification for deviations from standard practices. Additionally, case studies can help institutions contextualize their practices over time and explain why decisions or processes may have changed or evolved.

Documentation

Robust and accurate documentation of AI systems and related processes is a cornerstone of responsible AI efforts, particularly as they move from ad-hoc to institutionalized efforts. Documentation can facilitate knowledge-sharing across teams, anchor governance process, inform decision making, and support public communication and transparency efforts.[43]

- **Begin with the basics.** Documentation efforts should start with a basic accounting of what AI systems are in development or use. Then, practitioners can layer in additional details about systems, results of tests or assessments, mitigations or controls that have been applied to different models or systems, and decisions have been made about the conditions under which the system is permitted to launch or be used.

  A variety of documentation practices can serve responsible AI efforts, and the relevant details to include may continue to shift as governance processes mature. As a starting point, practitioners should identify a few core elements to document about all AI models or systems; this can be then used as a scaffold into which evolving documentation needs

[42] The full text of footnote 42 can be NIST AI 100-1: Artificial Intelligence Risk Management Framework (AI RMF 1.0). https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf [https://perma.cc/X2QH-7HKZ]

[43] Winecoff, A., & Bogen, M. (2024). Improving governance outcomes through AI documentation: Bridging theory and practice. Center for Democracy & Technology. https://cdt.org/insights/report-improving-governance-outcomes-through-ai-documentation-bridging-theory-and-practice/ [perma.cc/FN2H-GZSY]

can be integrated.

- **Tailor and structure documentation to meet a variety of governance goals.** The success of documentation in supporting internal risk management depends on whether documentation is appropriately tailored to the various needs of technical teams, risk management professionals, decisionmakers, and other stakeholders like policymakers and users. By regularly refining these artifacts and evaluating their efficacy, organizations can enhance collaboration, governance, and decision-making, ultimately contributing to improved AI system risk management over a long-term horizon.

- **Set up ways to share information thoughtfully across teams.** Sharing sensitive documentation about responsible AI issues, including about existing risks and AI use cases, can sometimes be challenging within an organization because of concerns about legal obligations and liability. To navigate these constraints, teams may need to determine how to share findings in more abstract ways to ensure legal compliance. This can require developing thoughtful communication channels, access controls to sensitive documents, and clear rules about what information is and is not allowed to be distributed.

- **Use documentation to facilitate internal legibility of responsible AI work, which can help avoid duplication of efforts and support consistency.** Before investing in new strategies or solutions, practitioners should first assess what existing practices are already in place within their organization to ensure that efforts are additive and not redundant. Building institutional knowledge that is both accessible and actionable can support consistent and efficient adoption of responsible AI practices; a robust repository of internal case studies of past projects can help teams to understand whether someone has already tackled a specific problem and what the results were. But without a clear process for surfacing and sharing insights, teams risk duplicating efforts or missing key lessons learned from past projects.[44] To capture the full benefits of documentation, practitioners should ensure the documents are easy to access and are known across the organization.

  "Most people won't be aware of the work we are doing (and will not look) unless we keep telling them."

[44] For a more detailed exploration of the role of documentation in enhancing AI governance, see Winecoff, A., & Bogen, M. (2024). Improving governance outcomes through AI documentation: Bridging theory and practice. Center for Democracy & Technology. https://cdt.org/insights/report-improving-governance-outcomes-through-ai-documentation-bridging-theory-and-practice/ [perma.cc/FN2H-GZSY]

Progress: Track efforts holistically

To maintain sufficient investment and enhance the impact of responsible AI work, its leaders and teams must define, measure, and show evidence of success — both internally and externally. Teams should establish effective metrics for assessing the performance and impact of specific activities, develop maturity models to evaluate the robustness of their practice, and commit to public transparency.

Metrics

Measuring and demonstrating progress is crucial for practitioners to track success and areas for improvement, show that resources are being used effectively so they are not reallocated to other company projects that may seem to be more pressing, and hold organizations accountable for claims they may make about their investments in responsible practices. This is hard, but critical, to do well.

- **Establish metrics to evaluate success, even if it seems challenging.**[45] Setting good metrics is especially important given the qualitative, cross-functional nature of responsible AI. Many organizations have established mechanisms for progress and performance measurement that tend to rely on simple heuristics and countable objects; for example, products shipped, hours billed, sales made, deals closed, and grants won. These metrics are not always possible to use — or desirable — for responsible AI work.

  At the same time, practitioners must weigh demands for quantification to show progress with the risks of oversimplifying responsible AI work, since overly quantitative metrics can obscure the reality on the ground or fall prey to Goodhart's Law ("when a measure becomes the target, it ceases to be a good measure").

  Finding the right balance between simplification for the sake of quantification, on the one hand, and capturing a complex reality for the sake of accuracy, on the other, will depend on the kind of organization and what the metrics are being used for. For example, for resourcing requests, some sort of quantifiable metrics may be necessary so that internal leaders can more effectively compare requests alongside those made by other teams; for communicating impact of the work to external audiences, qualitative narratives may be more important for promoting transparency and maintaining credibility. Notably, merely setting up a team or process is not a metric of success. Success requires teams to do tangible work that has a real impact, and pointing to the existence of teams without backing them with resources or empowering them to

[45] Madaio, M., Egede, L., Subramonyam, H., Vaughan, J. W., & Wallach, H. (2022). Assessing the fairness of AI systems: AI practitioners' processes, challenges, and needs for support. ACM on Human-Computer Interaction. https://doi.org/10.1145/3512899 [perma.cc/4X7N-MRM4]

implement responsible practices is a prime example of ethics washing.

- **Use metrics as a tool to drive responsible AI efforts.** Responsible AI work is complex and evolving, so developing clear metrics can be important to help practitioners understand how to set goals and feel they are making concrete progress in a rapidly changing field. On the other hand, because responsible AI is highly context-dependent, teams should take care when establishing quantitative metrics. One approach practitioners can consider is thinking about "coverage," or the number or proportion of teams or products that have adapted their procedures to include a set of responsible practices. For instance, using maturity models (discussed below), teams can be guided to adopt a series of increasingly robust practices over time, and tracking coverage of which teams have moved through which stages of such a framework can help organizations understand the trajectory of the organization's overall responsible AI efforts.

- **Measure and demonstrate success to maintain team morale, retain talent, and develop a community of practitioners motivated to do RAI work.** According to practitioners, technologists can become frustrated when they think "they're doing good work and then are told it's not enough." And if practitioners are situated in product or research teams where all staff are judged on a predefined rubric, failure to translate responsible AI milestones to fit this rubric can mean that skilled personnel are not recognized — or are even penalized — even when they are making significant contributions to risk management efforts. Even with initial leadership buy-in, if teams are not set up to demonstrate progress, leadership support may erode over time or fall victim to destabilizing reorganizations that can slow the institutionalization of foundational risk management efforts.

  "Working in RAI can often feel like chasing the horizon and you don't have as many feelings of accomplishment like launches and milestones as you do on traditional product work."

Maturity Models

A maturity model is a framework for assessing the progress of an organization's capabilities; in the context of responsible AI, they can be used to assess organizational maturity regarding the integration of responsible practices, and can help organizations better understand how their responsible AI efforts are shaping up.

- **Assess your organization's responsible AI capabilities against a range of factors — culture, organizational structure, and methodologies — using a maturity model.** Maturity models can focus on organizational capacity or adoption, or on sophistication of responsible AI methods themselves. Although terminology may vary, the common structure for maturity models includes: beginner or basic, intermediate or emerging, advanced, and leading.[46]

In the beginner stage, a responsible AI practice is either non-existent or latent. The organization likely has not articulated a scope for responsible AI work or developed and empowered practitioners to undertake and develop a responsible AI practice. Perhaps a responsible AI team has been established, but the substance of its work is still immature. In intermediate stages, the foundations for responsible AI work have been established; there may be, for example, a dedicated team with enough teeth and resourcing to accomplish key objectives; responsible AI is embedded in the product development lifecycle and a key part of organizational decision-making; and important tooling and processes, from gap assessment to documentation to model inventories, have been introduced. In advanced stages, a responsible AI practice is integrated into the organization's structure, culture, and leadership, and responsible AI work is rigorous and meets high industry standards. Finally, in the leading stage, a responsible AI practice is at the vanguard in developing standards and tools, and is likely helping inform and support industry-wide practices and even policymaking on AI.

- **Use both qualitative and quantitative methods to conduct the maturity model assessment.** Assessment of maturity can include organization-wide surveys, similar to those conducted by many organizations to assess characteristics of diversity or culture.[47] Other examples include semi-structured interviews or focus groups with representative cohorts from across the organization; while these are not anonymous, they have the advantage of allowing for deeper exploration of a particular issue that may not be captured in a survey question. More generally, maturity assessments foster a culture of evidence-based decision-making and continuous improvement, and encourage collaboration across teams.

  If formal assessments are not feasible, informal methods such as employee feedback sessions might be useful, particularly in cases where senior leadership are hesitant to collect data on sensitive use cases not required by law, or in newer organizations that lack the resourcing to conduct a more robust maturity assessment.

- **Don't mistake progress against maturity model benchmarks as evidence that AI-related risks are truly being addressed.** Meeting a particular maturity level is not in itself indicative that there are no ethical or governance concerns with the development

---

[46] For a different example, see, CSIRO. RAI Maturity Model. https://research.csiro.au/ss/science/projects/responsible-ai-pattern-catalogue/rai-maturity-model/ [https://perma.cc/5JVS-7XTR]; or Vorvoreanu, M., Heger, A., Passi, S., Dhanorkar, S., Kahn, Z., & Wang, R. (2023). Responsible AI maturity model. Microsoft Research. https://www.microsoft.com/en-us/research/publication/responsible-ai-maturity-model [https://perma.cc/LL2L-MUQL]

[47] Dotan, R., Blili-Hamelin, B., Madhavan, R., Matthews, J., & Scarpino, J. (2024). Evolving AI risk management: A maturity model based on the NIST AI risk management framework. arXiv. https://arxiv.org/pdf/2401.15229 [perma.cc/4X36-9TG5]

or deployment of AI in that organization; indeed, many organizations considered as 'advanced' or 'leading' have engaged in product launches and internal practices that have raised concerns. Nevertheless, maturity model-based efforts can help organize disparate efforts across distributed organizations and encourage investment and prioritization even as best practices are still evolving — a critical consideration for AI governance work for which research is still ongoing and for which standards have not yet crystallized.

Transparent Public Communication

Civic organizations, the media, and the general public all have a stake in how AI is built, designed, and used[48] so providing public information clearly is key to inviting and informing civic discourse on AI ethics and governance as well as creating channels for feedback and improvement for organizations. In some cases, such transparency can even help identify and prevent harms from AI systems before they happen.[49] And for use of AI systems in the public sector, considerations of transparent public communications may be related to the public's right of access to information and other fundamental rights.[50]

To help enable stakeholders to track progress in responsible AI, companies and organizations should engage in public reporting while avoiding performative or opaque communications.

- **Provide regular updates on progress, even when explicit requirements are absent.** Organizations seeking to share progress on responsible AI with external stakeholders should consider providing regular, detailed updates, such as through quarterly or annual reports and updates on the organization's website. The information provided can help stakeholders understand how responsible AI practices are being implemented; in an ideal case, stakeholders would also have an opportunity to provide feedback and engage further.

- **Ensure communication is not performative or misleading.** Public communications and reporting on responsible AI should strive to inform, update, and educate on progress without misleading or confusing audiences (e.g., by selectively showcasing responsible

[48] Alexander Buhmann and Christian Fieseler, Towards a deliberative framework for responsible innovation in artificial intelligence, Technology in Society, February 2021, https://www.sciencedirect.com/science/article/pii/S0160791X20312781

[49] Turner Lee, N. (2023). Making AI more explainable to protect the public from individual and community harms. Brookings Institution. https://www.brookings.edu/articles/making-ai-more-explainable-to-protect-the-public-from-individual-and-community-harms/ [perma.cc/T7KD-J8FL]

[50] Global Partnership on Artificial Intelligence. (2024). Algorithmic Transparency in the Public Sector. https://wp.oecd.ai/app/uploads/2024/12/14-Algorithmic-Transparency-in-the-Public-Sector-A-state-of-the-art-report-of-algorithmic-transparency-instruments.pdf [perma.cc/9PBG-QE8X]

AI artifacts).[51] Superficial engagement can pose reputational risks, leading stakeholders to question the integrity of the responsible AI practice or wonder what the organization is trying to hide. Conversely, transparent, well-explained, and clear information can have the added benefit of bolstering an organization's reputation. Crucially, transparency efforts should not lead to the offloading of duties for responsible innovation to end users, and must be set up to enable change.[52]

Conclusion

We have introduced the foundations of a responsible AI practice, articulated how practitioners can instantiate them in their organizations, and evaluated a range of associated considerations and tradeoffs. These are not, however, static decisions — responsibility is not a fixed destination but a continuous, evolving practice. Practitioners will need to adapt to a changing business, policy, and technological landscape. And while the external incentives for organizations to strive for responsible AI practices may ebb and flow, the challenges of developing and deploying AI at scale will remain; we hope the scaffolding provided in this report will serve those organizations eager to improve on the status quo.

As more and more companies become AI companies, practitioners and stakeholders will need to grapple with new, domain-specific challenges. The lessons we have captured from early and mature responsible AI practices will still, we think, resonate with those newer to the field. We welcome engagement and critical feedback, particularly as it relates to making responsible AI practices more concrete, tractable, actionable, and relevant for practitioners outside of technology companies.

In closing, we call attention to areas of investigation that require more attention: How can we ensure responsible AI is embedded not only in tools and processes but also in organizational decision-making cultures and incentives? What accountability mechanisms are most effective, especially when external regulation is evolving? How can responsible AI practices be adapted across organizations with vastly different sizes, capacities, and business models without losing rigor or relevance? These open questions reflect the challenges of a field that is still coalescing around norms, standards, and shared language, and highlight opportunities to learn from peers, refine organizational practices, and push for collective progress on the issue.

---

[51] Kawakami, A., Wilkinson, D., & Chouldechova, A. (2024). Do responsible AI artifacts advance stakeholder goals? Four key barriers perceived by legal and civil stakeholders. AAAI/ACM Conference on AI, Ethics, and Society. ACM. https://ojs.aaai.org/index.php/AIES/article/view/31669/33836 [perma.cc/U6K6-7N9Y]

[52] Kawakami, A., Wilkinson, D., & Chouldechova, A. (2024). Do responsible AI artifacts advance stakeholder goals? Four key barriers perceived by legal and civil stakeholders. AAAI/ACM Conference on AI, Ethics, and Society. ACM. https://ojs.aaai.org/index.php/AIES/article/view/31669/33836 [perma.cc/U6K6-7N9Y]

Find more about CDT's AI Governance Lab at *cdt.org*.

Acknowledgements
Thanks to the responsible AI practitioners who shared their insights with us as well as Drew Courtney and Samir Jain for their valuable feedback.

CENTER FOR DEMOCRACY & TECHNOLOGY

The Center for Democracy & Technology (CDT) is the leading nonpartisan, nonprofit organization fighting to advance civil rights and civil liberties in the digital age. We shape technology policy, governance, and design with a focus on equity and democratic values. Established in 1994, CDT has been a trusted advocate for digital rights since the earliest days of the internet.