

Business Roundtable Roadmap for Responsible Artificial Intelligence

Artificial intelligence (AI) is a powerful and versatile technology that will deliver transformational benefits to Americans' health, safety and prosperity. Fully unlocking AI's transformative potential will require supporting the innovative momentum around AI technologies and applications. It will also require building and maintaining public trust in AI — imperatives with shared responsibility across corporations and government.

All organizations engaging with AI should govern with purpose and innovate with integrity. Accordingly, Business Roundtable's *Roadmap for Responsible AI* articulates 10 core principles that businesses should consider in their journey toward Responsible AI.

These principles are underpinned by three foundational tenets:

1. First, a respect for the safety, dignity and autonomy of customers, end users and employees;
2. Second, a belief that concepts of effectiveness, fairness and security are inherent to Responsible AI; and
3. Third, an understanding that humans play a vital role across the AI lifecycle and that their involvement is at least as crucial as the trustworthiness of the technology itself.

Ten Core Principles

The Roadmap's 10 core principles are:

1. Innovate with and for diversity.
2. Mitigate the potential for unfair bias.
3. Design for and implement transparency, explainability and interpretability.
4. Invest in a future-ready AI workforce.
5. Evaluate and monitor model fitness and impact.
6. Manage data collection and data use responsibly.
7. Design and deploy secure AI systems.
8. Encourage a company-wide culture of Responsible AI.
9. Adapt existing governance structures to account for AI.
10. Operationalize AI governance throughout the whole organization.

The framing of these principles as a roadmap derives from four critical observations about the nature of the AI ecosystem:

- AI technologies and applications are dynamic, marked by continuous innovation by both developers and consumers. As such, AI is identified by its potential more than by any innate characteristics or today's technical methods and standards.
- AI technologies are not intrinsically good or bad. Companies design and use AI differently. Whether AI is responsible and trustworthy is best assessed through a contextual analysis of use and impact (outside of select industries or evolved practice areas within which risks are already specifically defined and regulated).
- AI risk is an important and necessary consideration, but it should be evaluated in comparison to the risk posed by existing human or technology alternatives.
- There is broad diversity in how and where companies in different industries engage along the AI lifecycle, from design and development to deployment and end use. There is also significant diversity across AI systems: Not all AI systems have direct implications for humans (for instance, considerations for AI deployed to optimize network performance will differ from AI used in employee recruitment applications).

These observations lead to two key conclusions:

There is no single approach or simple prescription for how to achieve Responsible AI. Rather, each company will activate these principles within the context of its own industry, use cases and workforce.

Most importantly, the individual and collective project of building and maintaining trust through Responsible AI is a continuous and evolving journey.

Trusted and Inclusive

Engage stakeholders to build trust in AI and equip them to participate in and benefit from AI innovations.

1. Innovate with and for diversity.

- Strive to assemble AI system design, development and deployment teams that represent a diversity of professional experience, subject matter expertise and lived experience — consistent with broader diversity goals and efforts.
- Apply similar diversity objectives to any cross-functional ethics boards, governance processes or oversight committees that address AI.
- Consider how and when diverse perspectives can supplement internal knowledge, particularly for high-consequence systems with direct human impact.

2. Mitigate the potential for unfair bias.

- Implement safeguards against unfair bias where AI systems may result in significant and high-consequence outcomes for individuals, while recognizing opportunities for AI to mitigate human bias.

3. Design for and implement transparency, explainability and interpretability.

- Explain the relationships between AI systems' inputs and outputs where possible and appropriate, particularly for systems that may result in significant and high-consequence outcomes for individuals.
- Tailor AI systems to the interests of different audiences, including deployers, end users and regulators as appropriate.

- Equip deployers of AI systems with sufficient information and training to support responsible and trustworthy downstream use.
- Disclose to end users when they are directly interacting with AI agents that simulate human interactions (e.g., chatbots).

4. Invest in a future-ready AI workforce.

- Determine which tasks AI will augment, change or eliminate, and consider where new tasks and jobs may be created.
- Support employees whose roles and responsibilities may change, and plan to reskill, upskill and/or provide new opportunities.
- Support education and skill-building efforts to accelerate the development of a modern and diverse workforce capable of developing and using AI responsibly.
- Make educational resources and training programs accessible across geographic, demographic and socio-economic backgrounds to broaden the AI talent pipeline.

Effective, Safe and Secure

Design, deploy and monitor AI systems to be fit for purpose and avoid foreseeable harms.

5. Evaluate and monitor model fitness and impact.

- Articulate clear goals and purposes for which AI models are optimized.
- Evaluate AI models in relation to specific use cases and contexts.
- Define performance metrics to capture the value as well as potential risk in AI design and deployment.
- Implement corresponding model testing, evaluation and training methods.
- Engage in self-assessments, whether such work is performed internally or against external guidelines or standards.
- Equip deployers of AI models with clear information about their proper functioning and limitations, including potential fitness for secondary purposes.
- Continually monitor and adjust models for fitness for purpose, accuracy and resilience.

6. Manage data collection and data use responsibly.

- Confirm that data is high quality, robust, appropriate for use and appropriately collected.
- Monitor and test datasets for accuracy and to avoid unfair bias, particularly for systems that may result in significant and high-consequence outcomes for individuals.
- Design data protection safeguards from the outset, collecting and using sensitive data based on demonstrated need.

7. Design and deploy secure AI systems.

- Account for the system's impact on security and whether it introduces or addresses security risks.
- Design and implement security safeguards from the outset to protect against adverse manipulation and preserve the system's intended function.

Accountable Governance

Implement effective governance to achieve benefits and mitigate risks across the AI lifecycle.

8. Encourage a companywide culture of Responsible AI.

- Provide boards and executive teams with appropriate visibility into high-level AI strategy and principles, and equip them to assess AI opportunities and risks.
- Engage representatives from different business units and disciplines so that decisionmaking is informed by multiple perspectives.
- Promote a culture of openness and learning so that developers test and think critically about the ways in which flaws, including implicit bias, may be reflected in data, methods or systems.
- Participate in opportunities to educate, share information and progress the public discourse.

9. Adapt existing governance structures to account for AI.

- Integrate AI into existing governance structures for risk management, data, compliance and business ethics as appropriate, and establish new AI-specific governance structures and processes as needed.
- Integrate social responsibility business objectives (e.g., sustainability, equity and inclusion) into AI governance guidelines and objectives.
- Evaluate whether existing company processes allow for receiving and responding to stakeholder feedback on impacts and interactions with AI systems.
- Periodically assess the efficacy of internal governance programs for Responsible AI.

10. Operationalize AI governance throughout the whole organization.

- Allocate sufficient resources, including personnel, time and budget, to operationalize Responsible AI principles.
- Clearly delineate roles, responsibilities and the chain of command to enhance transparency, accountability and decisionmaking.
- Make appropriately tailored trainings available to all relevant stakeholders within the organization to augment general AI understanding and fluency.