Responsible AI Materials Overview

Responsible Innovation Labs | Last updated 8/7/23

Executive Summary

This is a sampling of available Responsible AI resources. For example, the Responsible AI Index includes dozens of responsible AI tools, principles and frameworks (including from the largest companies (Microsoft, Google, Meta, etc.). Some are broad and generally applicable, others apply to narrow use cases or particular risks. The White House has shaped voluntary commitments and a blueprint for an AI Bill of Rights. The OECD has developed responsible AI principles. The NIST has a comprehensive AI risk management framework (among many other resources). The Business Roundtable has developed a roadmap and policy proposals. NGOs like the Partnership on AI, DAIR, Center for Democracy and Technology and others have deep, domain-specific resources on particular parts of the issue. The following is an overview of some of the most relevant Responsible AI resources, with brief descriptions of key points from each.

Government

U.S. Department of Commerce

*NIST: National Institute of Standards and Technology AI Risk Management Framework

- Most actionable and adaptable U.S. government framework; includes framing, measuring, managing, and governing risks; NIST's AI Resources Center supports implementation

The White House

*Voluntary Commitments: Ensuring Safe, Secure, and Trustworthy AI

- Outlines voluntary commitments made by Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI; includes red-teaming, security measures, watermarking, and more

Blueprint for an AI Bill of Rights (and Responsible Innovation Labs Guidance Memo)

- Takes stronger positions than Congress; focused on civil rights, equity and algorithmic systems (efficacy, privacy, and protection), similar to Trump administration; intended to coordinate the efforts of a diverse set of federal agencies around a core set of priorities

United States National Standards Strategy for Critical and Emerging Technology

U.S. Congress

Senator Schumer launches major effort to get ahead of artificial intelligence

Legislation introduced on AI sub-topics:

- AI and Biosecurity Risk (Markey, Budd)
- No Section 230 Immunity (Hawley, Blumenthal)
- *Privacy:* Children and Teens' Online Privacy Protection (COPPA 2.0) (Markey), Kids Online Safety (KOSA) (Blumenthal), No Robot Bots Bosses (Casey)
- *Defense:* "AI,Group of Four" Amendments in National Defense Authorization Act (Schumer, Rounds, Heinrich, Young), Block Nuclear Launch by Autonomous AI (Markey, Lieu, Beyer, Buck)
- *Investment:* Outbound Investment Transparency (Cornyn, Casey), Transparent Automated Governance (Peters), Expanding AI Research Access (Eshoo, McCaul, Beyer, Obernolte)

U.S. State Department

State Department perspective on artificial intelligence

U.S. State Department Guiding Principles on Government Use of Surveillance Technologies

U.S. Securities and Exchange Commission
Proposed Rules: Conflicts of Interest and Predictive Data Analytics | Full text here
- Focuses on practices of brokers-dealers and investment advisors

Federal Trade Commission
FTC: Joint Statement on Enforcement Efforts Against Discrimination and Bias in Automated Systems
- Note: The FTC's Commercial Surveillance and Data Security Rulemaking process started last year; new rules may be expected upon completion of this process

Federal Drug Administration
FDA releases two discussion papers to spur conversation about artificial intelligence and machine learning in drug development and manufacturing

European Union
EU AI Act: Proposal for Regulation laying down harmonised rules on artificial intelligence | In final stage
OpenAI reinstates service in Italy with enhanced transparency and rights for European users

China
Interim Measures for the Management of Generative Artificial Intelligence Services
- First-of-its-kind rules for generative AI services; U.S. State Dept. reported that China leads the US in 37 out of 44 key areas of AI; note: China plans to double its investment in AI by 2026, with most in new hardware solutions

International Governance Organizations
UN Human Rights Council Calls for AI Transparency | Full session reports
Hiroshima AI Process: The G7's New Effort to Harmonize AI Rules

Select Publications: Research & Policy Organizations

***Coalition for Health AI**: Blueprint for Trustworthy AI Implementation and Assurance for Healthcare
- CHAI is the first to apply the NIST Framework specifically to the healthcare industry
- Proposes tangible steps; most notably, registries and an oversight body

***Center for Security and Emerging Technology:** The Main Resource is Compute: A Survey of AI Researchers on the Importance of Compute
- Mixed study results underscore the need for more data about the importance of compute for teams developing AI, since it is the most readily available lever for the US government to use to grow or hamper the AI ecosystem in the US and abroad

***Partnership on AI**: Resource Library
- 50+ technology and media companies, universities, philanthropic orgs, and civil society orgs call on PAI to convene experts around specific topics
- Toolkits for builders of new tech, ex. Guidelines for AI and Shared Prosperity

Carnegie Endowment for International Peace: Reconciling the US Approach to AI
- Congress's efforts give agencies tools/knowledge they need about AI but none of it is binding; may set the stage for more binding regulation in the future but not clear what will pass
- Includes summary of key U.S. AI policy actions 2019-2023

MITRE: Evidence-based list of exploratory questions for artificial intelligence trust engineering

- Asserts current frameworks are competing and/or vague; focus on developing trustworthy AI toolkits for engineering teams; trust is critical for adoption

Center for Democracy and Technology: HIPAA explainer for start-ups, Individual Rights One-Pager

Social Science Research Network: Subjects and Stages of AI Dataset Development: A Framework for Dataset Accountability

Global Index on Responsible AI: Frameworks and Tools on Responsible AI

OECD: AI Principles Overview

Distributed AI Research Institute (DAIR): Publications

Berkman Klein Center: Principled AI: A map of ethical and rights-based approaches to principles for AI

International Standards Organization: IT — AI — Overview of ethical and societal concerns

- High-level overview of AI ethical and societal concerns for 168 member standards bodies with expectation that they "aim to ensure that professionals who design, develop, or deploy AI systems and applications or AI-based products or systems, recognize their unique position to exert influence on people, society, and the future of AI"

Sample Responsible AI Frameworks: Private Sector

Large companies have scoped Responsible AI (RAI) principles in varying detail – including in line with White House efforts. Microsoft and Meta primarily emphasize what they're doing internally to incorporate RAI into their own products (Transparency). Meta focuses on the need for regulation and steps they are taking to influence regulation. Google is differentiated by tools and research they release to advance RAI.

Microsoft: Responsible AI | **Focus: Governance, Transparency**

- Customer examples, and documentation for specific Microsoft products around (1) Fairness (2) Reliability & Safety (3) Privacy & Security (4) Inclusiveness (5) Transparency (6) Accountability
- Aim: Operationalize RAI across Microsoft by setting company-wide rules through central effort
- Underway: Tools on Azure ML to help customers and internal engineering teams adopt RAI

***Google:** Responsible AI Practices | **Focus: Release tools, Best practice**

- States directly which types of AI projects they will/will not pursue, evaluating across 7 principles
- Aim: Data and tool sharing, released 200+ research papers on RAI and technical resources

Meta: Facebook's 5 Pillars of Responsible AI | **Focus: Influence regulatory landscape, Transparency**

- RAI team ensures Meta's AI systems are designed and used responsibly across: (1) Privacy & Security (2) Fairness & Inclusion (3) Robustness & Safety (4) Transparency & Control (5) Accountability & Governance
- Aim: Demonstrate Meta actions and need for thoughtful regulation and collaborative research around AI best practices

Business Roundtable: Responsible AI Overview | **Focus: Build and maintain public trust in AI**

- 200+ CEO association from America's leading companies in every sector of the economy
- Published 10 high level, wide ranging core principles on AI to build and maintain public trust

Appendix

DETAILS ON NIST FRAMEWORK: Algorithms must be…

1. **<u>Useful</u>**
    - **Valid** with respect to accuracy, operability and meeting intended purpose and benefit
    - **Reliable**: The ability of any item (in this case, an AI model/tool) to perform its required function without failure, under stated conditions and over a defined time interval
        - Key facets of reliability include failure prevention (defined as "the termination of the ability of an item to perform a required function), workflow integration, and robustness under dataset shifts
    - **Testable**: The extent to which an algorithm's performance can be verified as satisfactory in terms of meeting *all* standards for trustworthy AI, including topics such as robustness, safety, bias mitigation, fairness, and equity in both development and evaluation.
        - Testability requires a strong contextual understanding of the model and its intended use, including *where, why,* and *how*
    - **Usable**: Denotes the quality of the user's experience (including effectiveness, efficiency, and satisfaction) of an algorithm's output
    - **Beneficial**: The ***benefit*** of an algorithm should be measured by the algorithm's impact on its intended outcomes (effectiveness) and overall health through its intended (and unintended) use, weighed against deleterious effects and risks.
2. **<u>Safe</u>**: Safe AI systems are ones in which "human life, health, property, or the environment" are not put at risk of harm
3. **<u>Accountable & Transparent</u>**
    - **Accountability** describes the responsibility of individuals involved in the development, deployment, and maintenance of AI systems to maintain auditability, minimize harm, report negative impact, and communicate design tradeoffs and opportunities for redress.
    - **Transparency** reflects the extent to which individuals interacting with an AI system or whose data are input into an AI system have access to information about that system and its outputs, regardless of whether they are aware that they are interacting with an AI
4. **<u>Explainable & Interpretable</u>**
    - **Explainability**: A representation of the mechanisms underlying AI systems' operation
    - **Interpretability:** Meaning of AI systems' output in the context of their designed functions
5. **<u>Fair - with harmful bias managed</u>**
    - **Bias:** Disparate performance or outcomes for selected groups defined by protected attributes such as race and ethnicity, and, in this paper, differences that are perpetuated and/or exacerbated by AI models and their use.
        - **Systemic bias** can be present in AI datasets; in the organizational norms, practices, and processes across the AI lifecycle; and throughout the broader society that uses AI systems
        - **Computational and statistical biases** may occur in datasets used to train AI systems and may also be present in the resulting algorithmic processes
        - **Human-cognitive biases**, as defined by NIST, are those that relate to how an individual or group perceives AI system information to make a decision or fill in missing information, or how humans think about the purposes and functions of an AI system
    - **Algorithmic fairness:** Multidisciplinary field of study that seeks to define, measure, and address fairness as it relates to algorithms used for decision- making. There are several key aspects to consider for algorithmic fairness: better design of new algorithms being built, audits of performance and consequences of currently used

algorithms; and examination of the consequences of algorithm use on a regular cadence

6. **Secure & Resilient**
    - **Resilient**: AI systems, as well as the ecosystems in which they are deployed, may be considered resilient if they are able to withstand unexpected adverse events or unexpected changes in their environment or use, or if they can maintain their functions and structure in the face of internal/external change, degrading safely and gracefully when necessary
    - **Secure**: Systems can maintain confidentiality, integrity, and availability thru protection mechanisms that prevent unauthorized access and use may be said to be **secure**
7. **Privacy-enhanced:** NIST's definition of **privacy** refers generally to the norms and practices that help to safeguard human autonomy, identity, and dignity