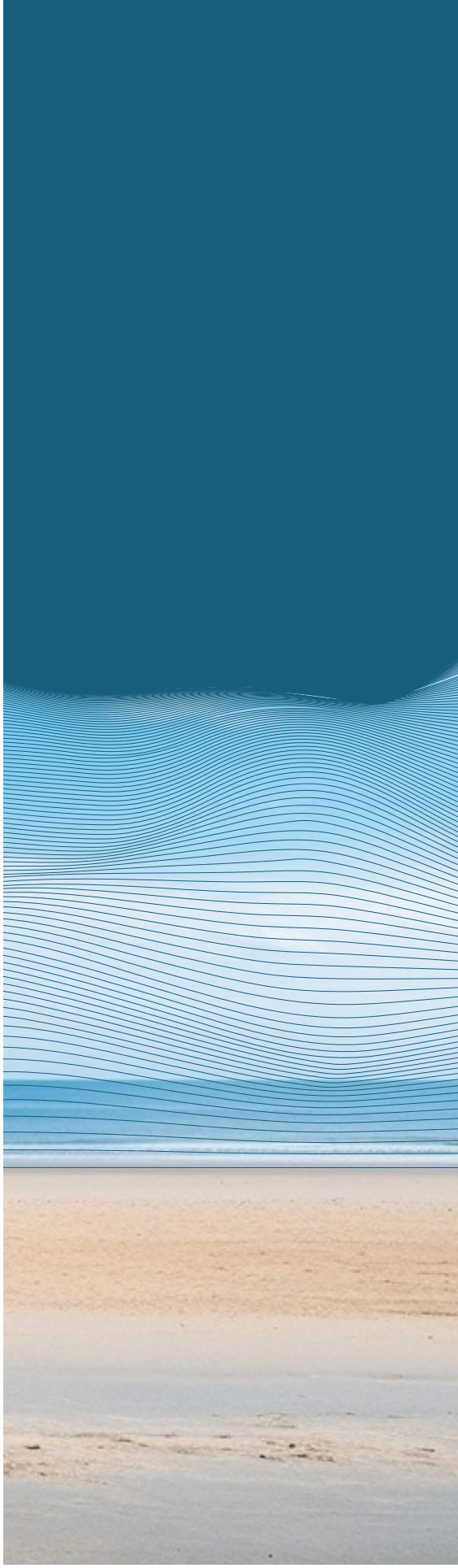
January 2025

Elder Research's [Responsible AI] Framework™

ELDER RESEARCH
DATA SCIENCE · AI · MACHINE LEARNING

Contents



Our Vision for Responsible AI at Elder Research

- Our goal is to lead and assist our clients in responsibly creating and maintaining AI[1] solutions.
- We aim to faithfully articulate how to practice the principles of Responsible AI (RAI).
- We will pursue opportunities to advocate for RAI in our industry by sharing our experience and contributing to thought leadership.

This document serves as the beginning of the next chapter of RAI at Elder Research. It outlines our vision, provides structure to how we think and talk about RAI, and helps us be deliberate in RAI consideration.

It is important to note that there will be more steps to come and further editions of this document as the field evolves. While we recognize the significance of future steps like formalizing tooling and processes, our focus here lies on setting the vision for what is ahead of us.

As we invite you to join us on this RAI journey, we emphasize that our aspirations drive us forward while recognizing the inevitability of change. Let's continue fostering Responsible AI practices at Elder Research, with our clients, and in our industry.

1 See Appendix, "AI Defined"

Why a Responsible AI Framework Is Needed

As AI applications become more widespread, organizations need to consider the ethical, legal, and reputational impact of their AI solutions. Governments around the world are attempting to address these same concerns with new regulations.

Establishing a Responsible AI framework enables developers and stakeholders to ensure their AI systems are not only effective but also ethical, trustworthy, and aligned with societal and organizational values. The natural outcome of this proactive approach is AI solutions that are more reliable and effective in the long term.

> "The natural outcome of this proactive approach is AI solutions that are more reliable and effective in the long term."

At Elder Research, this framework is essential for making Responsible AI clear and measurable in our diverse work—from client deliverables to thought leadership. By adhering to it, we will consistently deliver on our promises of reliability, safety, fairness, cooperation, and accountability and be prepared to comply with emerging laws and regulations.

How We Will Practice RAI Principles

We will faithfully articulate what Responsible AI (RAI) means and how it can be accomplished. RAI will become even more embedded in our principles as an organization—our professional conscience. Because we aspire for RAI to be integral to all our work, we will:

- Embed RAI principles into all our methodologies and toolsets
- Expand and adapt the principles to address generative AI challenges
- Articulate project-specific RAI considerations and plans with our clients
- Stay informed on RAI laws, regulations, and best practices
- Adopt standard definitions of RAI terms and principles
- Educate each other, our clients, our partners, and our industry on the practical application of RAI
- Stay true to our core value of integrity by choosing truth over expediency

At Elder Research our desire is to make RAI more than just a stand-alone service or a list of step-by-step requirements for technical delivery.

Instead, we want Responsible AI principles to be woven into the fabric of our work. By pausing to check that critical RAI questions have been asked and answered, we will make great strides toward that. We will work together to determine the best ways to record our answers to these questions, even when the answer is “not applicable,” or “not a priority.”

Our Diverse Scope of Influence

As a consultancy we strive to ensure the AI solutions we assess, develop, and deliver are responsible. The way we influence that may look different based on the situation. Sometimes it may look like providing advice to clients looking to implement AI solutions using cloud-based generative models. Because early RAI requirements tend to be marginal and subjective, as the solution development team, we desire to work with stakeholders and colleagues to fully define and refine requirements. Using this RAI framework as a guide, we want to help our clients harness the power of technology thoughtfully and responsibly.

Other times our influence may look like finding ways to apply RAI to a narrowly scoped technical service or while working with a limited time frame. In other instances it may come in the form of sharing lessons learned on our journey implementing RAI. Even when maintenance, governance, and ownership of an AI solution is not fully under our control, we aim to embody and model RAI principles in all our work.

> “... we aim to embody and model RAI principles in all our work”

We can think of this in terms of the nature of our responsibility in any given engagement, as articulated in this RASIC chart:

Responsible	In the performance of our own tasks, we faithfully adhere to RAI principles. We legibly and explicitly communicate the choices and reasoning related to RAI in our artifacts and deliverables.
Approve	We ask ourselves RAI questions to inform our approval decisions.
Support	We practice RAI principles in every supporting role.
Inform	We applaud RAI when we see it and consistently encourage efforts to improve.
Consult	We train, guide, and encourage our clients to embrace RAI principles and behaviors.

The Elder Research RAI Framework

Below is a summary of important aspects of a Responsible AI solution stated as goals. We recognize that perfection in RAI is not generally attainable; these goals are aspirational. They apply to all AI solutions, including generative AI.

Responsible AI analysis should assure that the RAI concepts and questions have been deliberately considered, rather than demanding perfect adherence. Openness and transparency are paramount and are prerequisites to continuous improvement.

The pyramid shape was chosen intentionally to display the hierarchy of these goals. They are not achieved independently but build on each other.

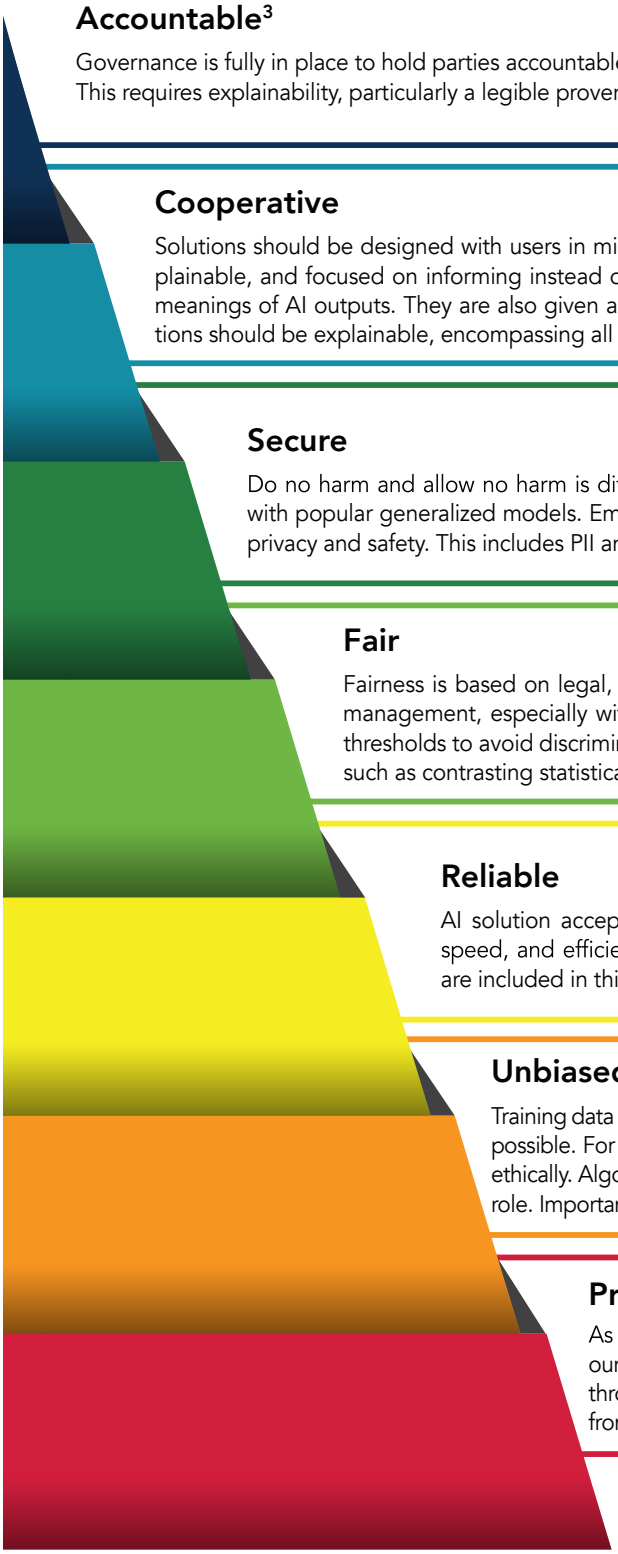
Without responsible, professional behavior, no other goal can be achieved. Without objective measures of bias and variance, fairness and reliability cannot be certified in any domain. Measurable reliability comes before security, and fairness can be quantifiably assured. Then we are prepared to take additional measures to cooperate with humans who need control of the solution and the distinct operational decisions the AI solution is designed to serve. Finally, with the foundational elements in place, including the provenance of AI solution outputs, the entire solution is governable, and accountability is clear.

The following page includes a description of the RAI goals outlined. Their definitions are not meant to be exhaustive but to be a guide on the types of consideration that should be made and in what order.

This framework deliberately consolidates near-synonyms common in public RAI frameworks.[2]

2 See Appendix, "Sources of Elder Research RAI Framework Goals"

Responsible AI Framework with Descriptions



Accountable[3]

Governance is fully in place to hold parties accountable for the solution, including its failures. This requires explainability, particularly a legible provenance.

Cooperative

Solutions should be designed with users in mind. This human-centric, cooperative approach means the solution is clear, explainable, and focused on informing instead of controlling choices. Users have clarity on the critical inputs, limitations, and meanings of AI outputs. They are also given appropriate rights and resources to override AI-recommended decisions. Solutions should be explainable, encompassing all transparency, traceability, and interpretability requirements.

Secure

Do no harm and allow no harm is difficult to achieve without full control over deployment and usage, particularly with popular generalized models. Emergency shutdown/fallback mechanisms and guardrails are in place to protect privacy and safety. This includes PII and IP and safeguards to respect human autonomy, identity, and dignity.

Fair

Fairness is based on legal, societal, and organizational values and should be audited case by case. Bias management, especially with model training data, is a prerequisite. One example is setting decisioning thresholds to avoid discrimination against protected classes of people. Fairness should be well articulated, such as contrasting statistical parity (zero bias) with equalized odds or equal outcomes.

Reliable

AI solution acceptably generalizes across time, place, and interest group, with performance, speed, and efficiency being quantified. The related terms of robust, performant, and accurate are included in this goal to be reliable.

Unbiased[4]

Training data correctly represent the scope of the implemented model (model footprint) as far as possible. For both prescriptive and generative AI, protected groups are represented fairly and ethically. Algorithmic bias (such as imposed by regularization or algorithm selection) may play a role. Important systematic deviation causing bias should be made legible and communicated.

Professional

As responsible professionals we only accept work that does not directly conflict with our values of integrity and respect. We commit to professional behavior of personnel throughout the lifecycle of an AI solution. This includes deliberately seeking reviews from outside eyes (reviews by people other than the creator).

3 See Appendix "Provenance of AI Outputs"
4 See Appendix, "Measuring Bias"

Core Questions: Business Leaders, Practitioners, and Users

To successfully implement the RAI framework, the perspectives of business leaders, technical practitioners, and users should all be considered. These three perspectives lead to the questions and activities needed to consistently deliver Responsible AI solutions.

The activities are aspirational, and our position to influence them varies (see section "Our Diverse Scope of Influence" above). Further, what can and should be done depends on the client and their use case. RAI decisions should be deliberate and legible and come through our client partnerships.

Business Leaders

We encourage business leaders (owners of AI solution) to write down clear answers to the questions below. As use cases and priorities change over time, businesses should regularly return to these questions to make necessary updates.

AI Goals

- What are the AI goals in terms of business/organizational outcomes?
- What are the fit-for-purpose objectives?
- Which people-driven systems and processes will be affected?
- Do any aspects of the AI solution risk violating the principles of integrity and respect?

Stakeholders

- Who are the stakeholders associated with the value chain?
- Who owns the value enhanced/created by the AI solution? (business owner)
- Who supports the platform(s) required by the AI solution?
- Whose jobs/tasks are impacted by the AI solution?
- Who governs decisions about deploying/using/refreshing/decommissioning the solution?
- Who will use the AI outputs to make or inform decisions?
- Who is subjected to decisions informed by the system?

Measures of Success

- What measures of value will be produced and what losses will be avoided?
- How will fairness be measured?
- How will compliance be measured?
- What measures will determine human compatibility and end-user acceptance?

Protection Requirements

- Essential Role of End User
  - Are the end users adequately informed about the model outputs, including its biases and limitations?
  - If a human is in the loop, is there a risk of automation bias (too readily accepting the model output without question)?
- Job Security Considerations
  - What processes will change or need different levels of human resources?
  - What retraining and reassignment is appropriate?
  - What is the employee communication plan?
- Fairness/Equity to Protected Groups
  - What regulations should be considered?
  - How do organizational values of fairness and equity apply?
- Consumer/Customer Privacy Requirements
  - What data is acceptable for training the model?
  - How might this cause leakage of PII?
- Safety Against Malicious Use
  - How can it be a tool for criminal actions?
  - What insider threats should be considered?
- Protecting Intellectual Property Rights
  - Does the solution leak IP?
  - Can the solution be stolen/reverse-engineered?
  - Are outputs protected against inappropriate use?

Solution Governance

- How frequently will measures be collected?
- How frequently will reports be created for each stakeholder group?
- What channels will be used to communicate with stakeholders?
- What are the triggers for action/intervention?

Technical Practitioners

As builders of AI solutions, we should consider the these questions at each stage of the product life cycle. We encourage our collaborators and those we influence to do the same.

Problem Formulation

- Is the purpose inherently unethical, disrespectful, or dishonest?
- Who is affected?
  - Remember rewards to one group over another should be defensible.
- Who are the regulatory stakeholders?
- Who are the public interest stakeholders?
- What are the feasible model update/refresh expectations?

Training Data Collection

- How well does training data reflect the target population?
- What vulnerable, smaller, or minority groups should be well represented?
  - This may require extra data gathering.
- Is privacy being honored?

Testing Strategy Development

- What special population groups are designated?
  - Partitioning by time period and special interest group, held out from training for testing, is likely appropriate.
- How will you test for acceptance criteria, including regulatory compliance?
- For ongoing model monitoring purposes, what special tests need to be done for regulatory or fairness requirements?

Model Specification

- What are the explainability requirements for data, model development process, and model structure? What assumptions/limitations of the model must be documented?
- What are the robustness requirements for model and system performance?
  - Are they associated with specific time periods or designated special groups?
  - Are there any special edge cases when they are out-of-sample?
- Is the solution compatible with all target platforms (for all user groups)?
- Is the scope for the solution (AI footprint) clearly communicated to avoid misapplication of the model?

Model Testing

- Does the testing environment mirror the production/operational target environment?
  - We like a controlled pre-production sandbox, analogous to clinical trials for pharmaceuticals.
- Scripts for Testing, Covering All Acceptance Criteria
  - Scope checking: Does the training and test data contain adequate data to represent the complete intended scope of application?
  - Random cross-validation: Is there unacceptable algorithmic overfitting? This is the first but minimum performance test.
  - Explicit groups: Has the model been tested by explicit groups, including protected classes and process-oriented time periods? For each group note the confidence in model performance.
  - Code Review: Has version control software been used? The solution may also require that code is released to production by an approver who is not the writer/requestor.
  - Adversarial testing: Does the solution require adversarial testing, akin to red teaming, including for edge cases?
- What alternative specifications have been considered and tested?
- Outside eyes: Have testing and reviews been conducted by person(s) other than the creator/developer(s)?

Model Refitting

- If a refitting cadence is anticipated, does this include regression testing (on old data) of each refit? What are the reporting requirements for these?

Model Deployment and Redeployment, Including Refit and Rebuild

- Is the provenance of solution outputs documented?
  - Describe/illustrate decisions, inputs, and tools (methods) to generate the solution and deliver outputs.
  - How will end users be informed about basis for AI output, including the data used, the information *not* used, and assumptions? This is critical for them to judge when to *not* accept the outputs.
- Assure all monitoring and testing scripts will run in production, including tests for scope compliance.
- What measures are in place to contain the footprint?
  - This is model management and governance.
  - On the access control side, make sure only those authorized can use the model.
  - On the model deployment side, build in ways to ensure RAI principles are not ignored.
- Are there adequate guards against data (analytic output) leakage to avoid harmful use of the model?
- Are there any insider threats to consider?
  - Can insiders use the output for personal gain?
  - Is the transmission of the analytic output tracked (who may and who did access it)?

Model Governance

- Are the following monitoring tools in place if required?
  - Regression test scripts
  - Solution footprint monitoring scripts: the scope of addressed entities; reports on scope and fairness questions
  - Changes in data distributions (inputs and outputs)
  - External feedback collection systems
  - Value creation metric reporting
- Are governance standards set up as needed by the organization?
  - Reporting to Stakeholders
    - Include internal, regulatory, and relevant/selected third parties
    - Report failures/issues to adversely affected groups
    - Value created for interested stakeholders
    - Model health: performance and fit for purpose
    - Risk tracking
  - Decisions to Consider
    - Modify target footprint
    - Decommission
    - Refit
    - Rebuild

End User Engagement and Protection

Because RAI is human-centric, it is critical that leaders consider the factors needed to effectively engage and protect end users.

Do the users have appropriate training, such as the following?

- End User Guide
  - Online help
  - Reference materials
- Task-Specific Training
  - Develop clear training resources appropriate for each user role.
  - Deliver timely training to all users to ensure the solution is used safely and effectively.

Have users and stakeholders been provided with appropriate knowledge of risks? Common examples:

- How often will an analytic output, such as a prediction, be wrong?
- How far off can users expect an estimate to be?
- What inputs most impacted the prediction, estimate, forecast, or output?
- What override options are available?
- On what basis might users override model outputs?
- Are users compelled to behave in ways that are dishonest or disrespectful?
- What other information might users collect that the model could not?
  - For Large Language Models (LLMs) and other generative AI, facilitate direct query of reliable source material.
  - For fraud investigators, communicate reasons for being tagged.
  - For event risk, communicate primary drivers and point out unavailable latent factors.

Does the solution provide conservative direction? Examples:

- Tune the thresholds for classification to balance the cost of false negatives vs. false positives.
- Gravitate toward status quo, depending on the paucity of evidence for each instance.
  - Build strong Bayesian priors.
  - Gravitate toward null hypothesis.
    - Require small p-value.
    - Require small p-value across time periods and other dimensions.

Does the solution capture and record when users override/ deviate from AI output recommendations?

- Sense the overrides soon enough to communicate corresponding risk.
- Capture data relevant to the overrides.
  - Determine the who, what, when and why.
  - Report the resulting outcome for monitoring purposes.

Special Risks for Generalized Models

The safety and ethical requirements for generalized models[5] including foundation models and LLMs, can be quite intractable because human behavior is being mimicked. Guaranteeing a particular model virtue or behavior may be impossible in the same sense it is impossible for humans. Thus, the requirements for a generalized model look much more like a job description than a technical specification for software.

> "Responsible AI should deal with the scope of use of generalized models, not just the specification and training of the model."

Also, because such models perform a countless variety of tasks, the scope of potential users can be practically unbounded if access is not tightly controlled. Responsible AI should deal with the scope of use of generalized models, not just the specification and training of the model.

Consider the analogy of water as a resource. Water has an endless number of uses. It is widely available and is generally used for human benefit. But it can be wasted, a person can drown in a pool, and a single bomb could destroy a dam killing tens of thousands of people. Laws apply but cannot prevent its use, and the laws are never uniformly enforced. Criminals and bad actors find ways to use it maliciously. However, individual organizations can govern its usage responsibly—if not perfectly.

Also, generalized AI solutions are currently in an exponential growth phase. Societies will adapt and influence the solution types they accept and embrace. ***Rules should evolve with needs and public sentiment.***

Types of Generalized Model Risks

The motivations and users of a generalized model[5] are highly diverse and difficult to anticipate. Our attempt here divides generalized model risks into two groups: non-intentional and intentional (malicious). As time goes on, we expect other risks will become evident.

For now we should consider these risks whenever serving a role related to AI solutions.

5 See Appendix, "Generalized Model Definition"

Non-Malicious Intent Examples

- Hallucinations
  - Extrapolating carelessly from poor evidence to false statements of facts
- False Attributions
  - Pretending results were generated by humans
  - Crediting wrong party or no party for generated content
  - Misinterpreting statements
- Incomplete Information
  - Missing information too subtle to notice, but essential
  - Preponderance of conventional or institutional wisdom is not a guarantee of veracity
  - Example: highest ranked dermatologist, but some dermatologists are left out of scope
- Exposing Protected Information
  - Social Security numbers, passwords, keys, account numbers, calendars, and locations from training data
  - Tracked personal conditions and behaviors with privacy concerns
  - Leaked intellectual property
  - Implicit information from prompts by other users
  - Private company information

Malicious Examples

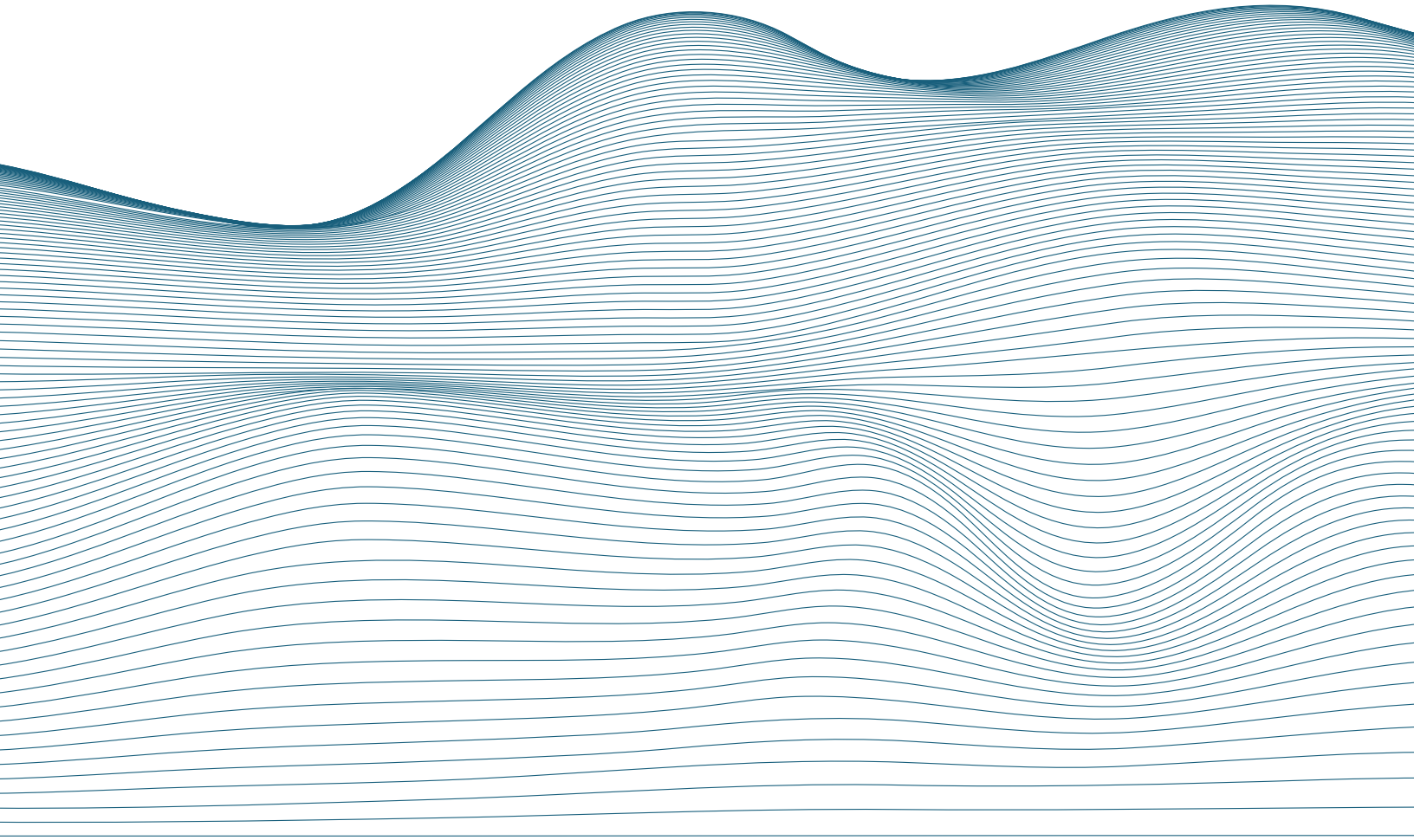
- Cheating on Skills Tests
  - Academic assessments
  - Job interviews
  - Sloppy lawyering, doctoring
- Blackmail
  - Maligning the reputation of another; producing false text, sound, images, or video with the intent to harm, such as to extort money
  - Efforts by developers of large language models to try not to produce inflammatory text are only marginally effective. Criminals can circumvent
  - Proving after the fact that the generated content is not real can be assisted by watermarking, up to and including the IP address that generated the image, for example.
  - We will be increasingly dependent on certifications of genuine content, probably in a registry of information resembling bank-level security.
  - A party can thereby certify the provenance of (hash of) a digital artifact—everything from TV broadcasts to deeds of title, copyrights, patents, biometrics, and affidavits.
    - For example, someone can produce an affidavit certifying they are the producers of a video and that it was *not* generated by AI.
- Social Engineering
  - For example, spear phishing to expertly persuade the disclosure of sensitive information
  - We need a recourse to check the provenance of a communication.
    - Did the email come from our domain?
    - Is there a traceable ID to investigate?
    - Is there a way to video call the person? If not, what preemptive actions should the organization take?

Governance and Accountability for Models on Closed Platforms

We encourage our clients to act responsibly and understand potential downstream harm of AI solutions. We encourage them to make sure each model is fit for purpose and that this purpose is in full compliance with their ethical, moral, and brand values.

> "... act responsibly and understand potential downstream harm of AI solutions."

When a client procures a foundational model for internal use, they generally groom its inputs and constrain its outputs. This means their organization likely retains legal accountability for harm produced by their AI solution—even though an outside model is a part of their solution.



Published RAI Frameworks and Policies

Public discourse and rule setting for Responsible AI are dynamic and constantly expanding. With widely varying formats and degrees of detail, many have published their frameworks. We are excited to add our framework to the conversation, distilling well-established RAI concepts and adding elements we consider critical.

A few notable Responsible AI resources are referenced below. While none of these resources match our needs exactly, Google's Responsible AI Practices reflects some of the ways we approach our work in RAI

Guides to Frameworks and Tools

- Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence: This White House statement shares guidelines on responsibly implementing AI.
- Awesome Artificial Intelligence Guidelines: Primarily for technicians, this broad GitHub roadmap outlines artificial intelligence processes, regulations, and more.
- A Matrix for Selecting Responsible AI Frameworks: The Center for Security and Emerging Technology provides RAI resources for different types of organizations and use cases.
- The Language of Trustworthy AI: This in-depth glossary from the National Institute of Standards and Technology (NIST) outlines key terms related to artificial intelligence.
- The Responsible Machine Learning Principles: The Institute for Ethical AI & Machine Learning provides easy-to-read principles of responsible ML development.
- Ethical Principles for Artificial Intelligence: The U.S. Department of Defense (DOD) shares its framework for ethical use of AI.

Global Governance Bodies Frameworks

- ISO/IEC 23894:2023 Guidance on AI Risk Management: This document provides guidance on how organizations that develop artificial intelligence solutions can effectively implement AI risk management.
- The Presidio Recommendations on Responsible Generative AI: Published by the World Economic Forum, this white paper provides 30 specific recommendations summarized into one paragraph each. Many recommendations apply to national and global governing bodies, but 10 of them apply to work Elder Research is often contracted to provide.
- Artificial Intelligence Risk Management Framework: This NIST framework broadly outlines the risks and recommended countermeasures.

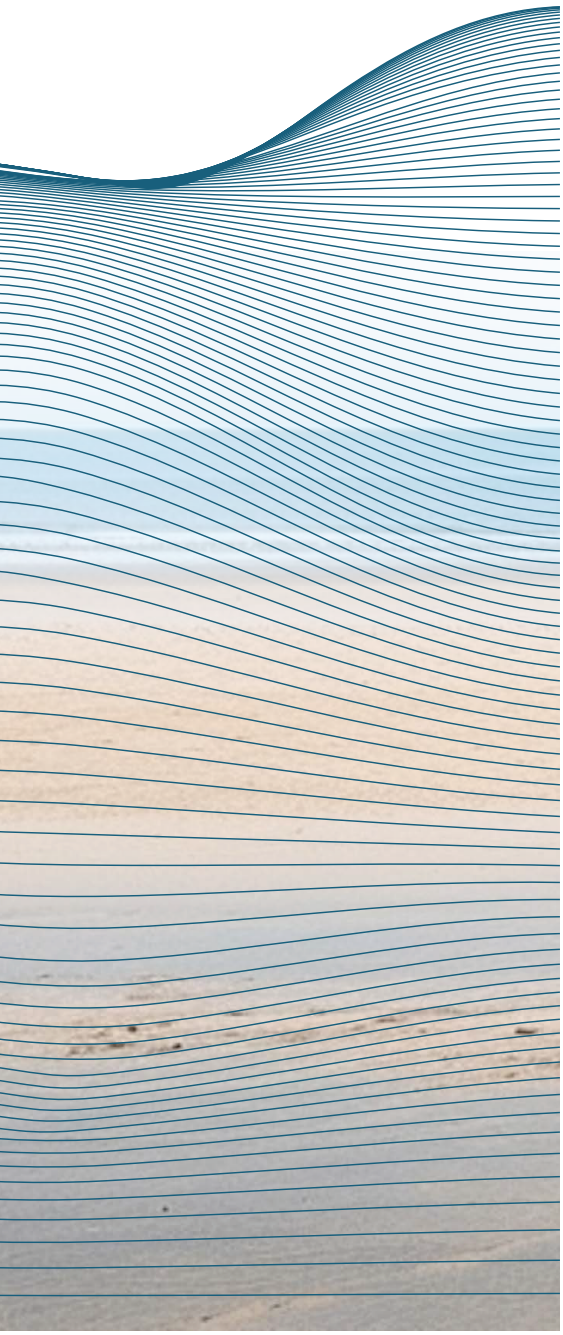
Corporate Frameworks and Advice

- Microsoft Responsible AI Standard, v2: This release outlines a prescriptive process for Microsoft developers.
- Responsible AI Practices: Google presents summaries of the principles and practices for RAI.
- Responsible Use of Machine Learning: Amazon gives broad guidelines for the major lifecycle stages of an ML system.

Consultancies

- Responsible AI Institute: This institute provides organizations with assessments and guidance on six major dimensions of RAI.
- PricewaterhouseCoopers: PwC has put together a survey and set of principles that abstract some of the key areas they've identified for Responsible AI.

Appendices



1. AI Defined

AI is defined here to include machine learning and artificial intelligence, common deliverables of our data science practice at Elder Research. NIST's Artificial Intelligence Risk Management Framework refers to an AI system as "an engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments."

AI has the meaning set forth in 15 U.S.C. 9401(3): "a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. Artificial intelligence systems use machine- and human-based inputs to perceive real and virtual environments; abstract such perceptions into models through analysis in an automated manner; and use model inference to formulate options for information or action."

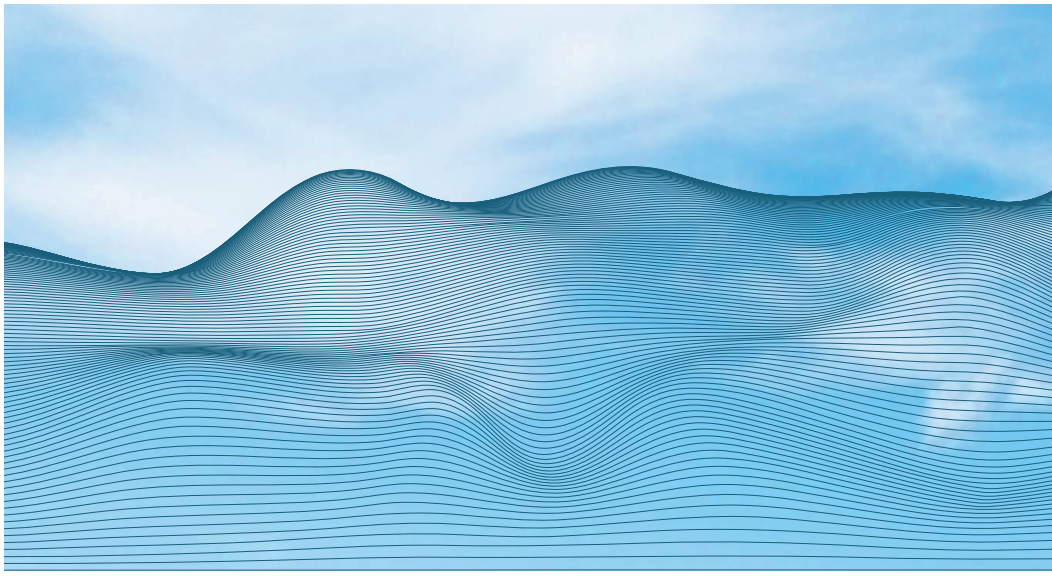
2. Sources of Elder Research RAI Framework Goals

Throughout all the publicly available recommendations, guidelines, and claimed best practices, there are no obvious conceptual disagreements on the concepts of RAI. But the public framework glossaries are not entirely uniform, some concepts are sometimes absent, and some are conflated. The DOD, for example, recognized this and proposed to resolve its terminology with NIST's glossary, with moderate success. See "Published RAI Frameworks and Policies."

3. Provenance of AI Outputs

A Responsible AI solution includes design documentation with a clear provenance of the solution's outputs. Adequate information about provenance makes accountability and governance possible.

1. What was the source of the training data for models within the AI solution? What were the processes that generated the data? Who holds responsibility for the completeness and accuracy of those data?
2. What model features from the data were used to train the model? How were rows (observations) selected for model training? What organization holds responsibility for training the model, given the data? What methodologies and policies do they adhere to?
3. What machine learning (ML) algorithm was selected, with its hyperparameters, and why? Who is responsible for the learning algorithm?
4. How do the features interact generally in the model to deliver the output? Meet the general transparency requirements for the solution.
5. How do the features interact specifically (for a given observation addressed by the solution) in the model to deliver the output? Meet the specific transparency requirements for the solution.



4. Measuring Bias

Bias, as defined here, refers to the calibration of model estimates. For example, an estimate is unbiased if the chances of the estimate being high are the same as the chances of it being low. Variance refers to the spread of the observed actual values around their estimates. We seek to minimize both bias and variance.

The first and most critical step in minimizing bias is data curation for model build and evaluation. The data should be representative of the data the model will be applied to in actual use. Or it should have reliable weights on each observation such that a weighted bootstrapped sample will be representative of the target population.

The following is a list of tests in (roughly) increasing orders of sophistication and assurance, given representative model build data. Different tests are applicable in different situations.

0 - Test on Train

When the trained model is applied to the training data, are the answers, on average, correct (unbiased)? This is listed as 0 because it provides only trivial assurance the model will work on new data in actual use.

1 - Random Partitioning

This is the first level of *out-of-sample* partitioning where the model is trained on a random sample of the model build data. The model is then applied to a complementary random sample of the model build data. Bias and variance are reported on this. This randomization can be done repeatedly to establish distributions of bias and variance.

2 - Time Period Partitioning

This is the most common test to emulate the real-world (vs. random) partitioning the model will encounter in practice. This assumes the model build data contains date timestamps of the time each observation was generated. This encounters the natural changes in the data generation process over time. The model build data is partitioned into sequential temporal time frames. Some (usually the last) time frame(s) are held out for testing, while earlier ones are used for model training/fitting. To get good distributions of bias and variance over time, a variety of sets of training and test partitions are used, for example leaving one time period out for testing at a time.

3 - Protected Group Testing

Fairness is an RAI goal, and fairness goals will include definitions of groups where fairness should be certified and measured. In all the previous tests (0-2), sub-partition the test partition by group, and again measure bias and variance. The distribution of bias should easily include zero, and standard deviation should be quite uniform across groups when divided by the square root of the number of observations in that sub-partition.

4 - Special Group Partitioning

To meet fit-for-purpose objectives, partitioning by other groups of interest may be appropriate, again assessing the distributions of bias and variance.

5. Generalized Model Definition

Generalized models as used in this framework are generative models as well as any models that have a very broad range of uses. LLMs fall into this category as well as systems that generate images, video, and complex audio such as music or voice mimicking. Multi-modal models are already a reality. Generalized models have practically countless uses.

A subset of generalized models is *dual-use* models as defined by the DOD: "The term 'dual-use foundation model' means an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters, such as by:

i. substantially lowering the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons;

ii. enabling powerful offensive cyber operations through automated vulnerability discovery and exploitation against a wide range of potential targets of cyber attacks; or

iii. permitting the evasion of human control or oversight through means of deception or obfuscation.

Models meet this definition even if they are provided to end users with technical safeguards that attempt to prevent users from taking advantage of the relevant unsafe capabilities."