WHITEPAPER

Embracing Responsible AI

Building trust, ethical use, and fairness in a dynamic business landscape

fractal

Introduction

As new AI technologies emerge, evolve, and expand in the ever-changing industry landscape—we have only just begun to scratch the surface of how they can be implemented and leveraged to improve our lives.

But, as with any technology, AI can be applied constructively and destructively, unpacking a distinctive set of challenges. With every new advancement, concerns such as discriminatory behavior and unintended biases against protected features, violations of privacy regulations, limited consumer trust, and legal implications have come to the fore. Refraining from considering these challenges is no longer an option.

As we approach the exciting opportunities AI creates one use case at a time; we must carefully examine its darker side. To do this, we need to ask ourselves some difficult questions:

Why Explainable AI in business decision-making?

As businesses embrace the power of AI, it becomes increasingly evident that we must also uphold ethical, inclusive, and responsible AI practices. These practices are paramount in ensuring the success and sustainability of our AI-driven future. At the same time, we must address the emerging risks associated with the growing utilization of generative AI (GAI) techniques, including hallucinations and over-reliance. Acknowledging and proactively responding to the critical requirement for RAI can exponentially influence risk management, underwriting services, customer contentment, and business profitability.

Encouraging a culture of ethical awareness and action

As new AI technologies emerge, evolve, and expand in the ever-changing industry landscape—we have only just begun to scratch the surface of how they can be implemented and leveraged to improve our lives.

But, as with any technology, AI can be applied constructively and destructively, unpacking a distinctive set of challenges. With every new advancement, concerns such as discriminatory behavior and unintended biases against protected features, violations of privacy regulations, limited consumer trust, and legal implications have come to the fore. Refraining from considering these challenges is no longer an option.

As we approach the exciting opportunities AI creates one use case at a time; we must carefully examine its darker side. To do this, we need to ask ourselves some difficult questions:

Responsible AI Framework

Principles →	Behaviors →	Enablers
Social well-being & planet-inclusive Cost-benefit analyses for net positive on humans, society and planet	**Contestability** Newspaper test	**Codebase** Plug-ins/Cloud agnostic APIs
Privacy & Safety Respect and protection from misuse	**Human-Centricity** Evaluate biases	**Chief RAI** RAI diagnostic & Committees
Fairness & Equity Ensure equitable outcomes	**Adaptability** Augment signals and update	**Education** Certifications
Robustness & Stability Reliable, repeatable outcomes	**Upskilling - Future of Work** Unlearning, relearning positively	**Adoption** Nudge toolkit
Accountability Humans responsible/governance	**Explainability** Decision factors traceable	**ESGdiagnostic** Circular Design
Transparency All deserve white-box decision choices	**Attribution** Copyright infringement and monetization	**Synthetic Data** Privacy by design

Principles of Responsible AI

Understanding the underlying principles behind RAI is the first step towards implementing frameworks, toolkits, and processes designed to address the potential negative impact of AI.

1. SOCIAL WELL-BEING AND PLANET-INCLUSIVITY

Social well-being can be negatively affected by AI through various means. For example, there may be concerns about job displacement and the potential adverse impact on employment rates and income equality.

AI systems also often require substantial computational resources, creating a considerable ecological footprint. The manufacturing and disposal of AI-related hardware can also generate electronic waste, adding to environmental concerns.

POTENTIAL SOLUTION

There is a consensus that AI has the potential to replace people who don't adopt its use. To tackle this, prioritizing an extensive and accelerated training program is crucial to upskill and cross-skill the most vulnerable workforce.

To mitigate the ecological consequences of artificial intelligence, we must prioritize reducing computational demands by optimizing the model architecture and training process. By employing efficient techniques, we can decrease energy consumption and carbon emissions.

2. PRIVACY AND SAFETY

AI systems are susceptible to vulnerabilities like data leakage and attacks, and GAI poses unique risks due to limited regulation. These vulnerabilities can lead to privacy breaches, the creation of harmful content, and other potential privacy, safety, and security concerns.

POTENTIAL SOLUTION

Comprehensive data governance, privacy by design, and Privacy-Enhancing Technologies (PETs) are vital for addressing privacy concerns in the AI workflow. Adopting a privacy-by-design approach integrates privacy considerations from the start, while regular privacy impact assessments ensure compliance. Robust security measures, for example, encryption, access controls, and security audits, also help safeguard the data used in AI models.

3. FAIRNESS AND EQUITY

Biases exist not only in the data we use but also in how we interpret that data. For instance, GAI models may exhibit stereotypes, such as generating images of only female nurses when prompted for images of nurses in a hospital setting. Fairness and bias in AI, including generative AI, can stem from various factors:

Biased training data can perpetuate biases in generated outputs. When the training data lacks diversity and fairness, the AI model may exhibit biased behavior, falling short of representing the intended diverse population it aims to serve.

Biases can be introduced through design choices made during AI algorithm development. Decisions on features, model architecture, or optimization objectives may inadvertently favor certain groups or attributes over others.

Lack of diversity and inclusion within AI development teams can contribute to biased outcomes. Incorporating diverse perspectives and experiences is vital for identifying and addressing biases effectively.

Hidden stereotypical and biased patterns in the underlying algorithm can be learned from the data, often unnoticed during human analysis. Additionally, correlations with personally identifiable information (PII) and proxies can inadvertently amplify bias during model training.

POTENTIAL SOLUTION

To address biases effectively, we must promote diverse and representative training data, establish clear guidelines for algorithmic design, incorporate fairness metrics, foster inclusive development teams, and continuously monitor AI systems.

4. ROBUSTNESS AND STABILITY

Maintaining consistency becomes a significant challenge when working with intricate and diverse data sets. Ensuring consistent patterns and distributions across various samples can prove challenging. The practical application of generative models can be limited by their unreliability in producing unrealistic, implausible, or inadequate outputs that do not capture desired attributes. Moreover, generative models can produce outputs incorporating information or features absent in the training data, leading to hallucinations. For instance, the model can produce varying responses to the same prompt concurrently or at different intervals.

POTENTIAL SOLUTION

Consistent and dependable GAI outputs require the implementation of strict quality control measures. This includes thorough testing, validation, and evaluation against predefined criteria and benchmarks. Robust testing should cover adversarial attacks, input variations, and edge cases. By addressing vulnerabilities through these measures, the reliability of GAI systems can be improved.

5. ACCOUNTABILITY

The development of AI requires shared responsibility among all stakeholders who are involved. Data scientists, engineers, researchers, designers, policymakers, and decision-makers each play a pivotal role in ensuring ethical and responsible AI practices.

POTENTIAL SOLUTION

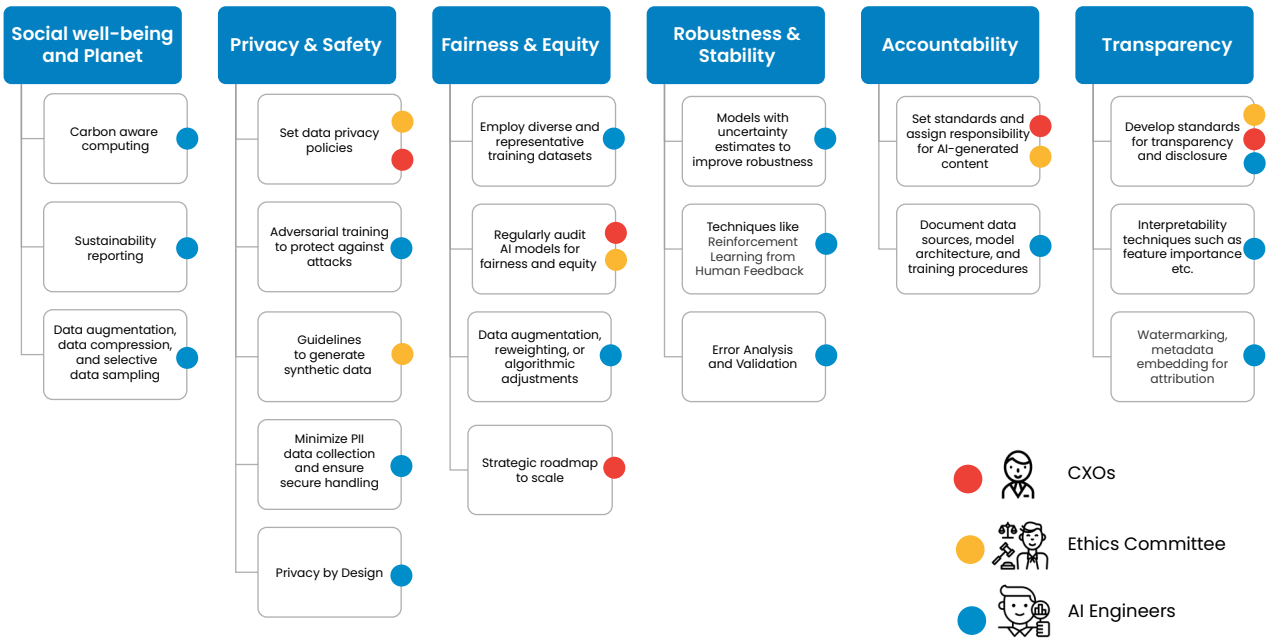
It is essential to establish a clear process that designates accountability for any issues arising from AI processes, minimizing potential challenges that may arise from multiple stakeholders. It is crucial to conscientiously address biases, errors, transparency, and user experience throughout the development of AI models while simultaneously adhering to ethical standards and legal regulations.

6. TRANSPARENCY

The difficulty in achieving transparency in AI systems arises from the intricate underlying mechanisms that produce their outputs, presenting a challenge of explainability too. The utilization of large data sets in the training process introduces complexities and biases that complicate the interpretation of their outputs.

POTENTIAL SOLUTION

To ensure transparency in AI solutions, we need methods to explain black box models and understand the reasoning behind AI outcomes. Sharing information about the model's architecture, parameters, and training data is vital. Integrating explainability methods into decision-making yields deep insights that improve transparency and comprehension.



The Dynamic Duo: Synergy of human behavior and AI systems

After laying out the fundamental principles, attention should be directed toward the actions of humans and AI systems. When considering behaviors, areas for consideration include the following:

1. CONTESTABILITY

Contestability is the ability for individuals and stakeholders to challenge or contest the decisions, processes, and outcomes generated by AI systems. For instance, if you applied for a loan and your application got rejected, it is important to have the right to contest that decision. The RAI framework should promote contestability as a desirable behavior. Users should be free to raise concerns or contest decisions related to ethical challenges or sensitive topics. Empowering users to contest AI outcomes is crucial for maintaining transparency and fairness.

2. HUMAN-CONCENTRICITY

When it comes to enterprise adoption of AI systems, the goal should not be to implement them for the sake of doing so. Instead, careful consideration and strategic planning are essential to ensure that implementing AI benefits your business. In today's tech-savvy world, users have become overly reliant on technological artifacts, making decisions primarily influenced by technological developments. Henceforth, it becomes crucial to prioritize human-centric AI systems, catering to their needs as the mainstay, rather than expecting individuals to put AI at the center.

3. ADAPTABILITY

Adaptability is another crucial aspect of RAI and requires the development of systems that can effectively respond to evolving data patterns, user requirements, and emerging ethical considerations. Businesses that keep abreast of shifting cultural and societal dynamics and technological landscape are well-positioned to tackle obstacles head-on proactively. This enables them to adapt their AI systems to align with evolving ethical standards and regulatory requirements, ensuring responsible and sustainable AI practices.

4. UPSKILLING

This involves acquiring new knowledge and skills about RAI principles, ethical decision-making, bias mitigation, and transparency. It entails understanding the societal impact of AI and effectively navigating the emerging ethical challenges. Through upskilling, individuals actively contribute to shaping RAI practices, ensuring that AI technologies are developed and deployed in a manner that aligns with ethical standards and upholds human values.

5. EXPLAINABILITY

It is crucial to prioritize the explainability of decisions made by AI systems. This entails establishing a robust framework that enables thorough auditing and traceability of the decision-making process. By doing so, we can provide a comprehensive and comprehensible explanation of the factors involved in making specific decisions and their underlying rationale.

6. ATTRIBUTION

The advent of AI has led to the generation of extensive volumes of new data, giving rise to copyright concerns as human-inspired content is reused creatively. The absence of appropriate attribution can raise queries regarding ownership, recognition, and compensation for the original creators of AI-generated content. For instance, Getty Images has initiated legal action against the creators of the AI art tool Stable Diffusion for unauthorized content scraping. The lawsuit alleges that GAI art tools violate copyright laws by extracting artists' work from the web without their consent.

Applying RAI in the real world — a use case example

Data drift refers to the changes in the statistical properties of training data over time, leading to a decline in the model's performance. In AI systems, maintaining reliability, fairness, and ethical use is vital - and that means tackling data drift head-on. Reasons for drift may include shifts in data distribution, alterations in variable relationships, or changes in environmental conditions.

Detecting and addressing data drift is essential for maintaining transparency, accountability, and fairness in AI decision-making. RAI practices involve regular data monitoring, ongoing model evaluation, and taking proactive measures to adjust and retrain models as needed.

Explainable AI (XAI) enhances transparency and trust by allowing humans to understand AI decision-making. XAI identifies biases/errors and generates human-readable explanations. It addresses data drift challenges, ensuring reliable and accountable AI systems.

Sales and revenue management for a CPG client

OBJECTIVE:

The objective is to enhance pricing and promotional strategies and uncover growth opportunities by analyzing sales and revenue data.

SOLUTION APPROACH:

Our proposed solution involves the utilization of RAI to analyze data effectively and detect any concept drifts. This allows for identifying patterns and trends in consumer behavior, ultimately enabling forecasting future sales by considering various factors.

HOW RAI HELPS:

RAI leverages its capabilities to analyze large volumes of data, detect concept drift over time, and continuously monitor data to ensure accurate and reliable sales projections. It further assists in optimizing pricing and promotional strategies.

IMPACT:

Implementing RAI enables informed decision-making, increasing revenue and profitability. It also helps identify emerging trends and consumer preferences, fostering a more data-driven and responsible business model.

Fractal's enablers

Codebase
Scalable custom codes with user-friendly APIs, full documentation, and endless customization options.

Chief RAI
Empowering RAI adoption with a taskforce for unified organizational integration and auditability

Education
Comprehensive RAI education for all levels, disciplines, and domains – technical and strategic

ICON

Adoption
Empowering developers with actionable checklists for ethical, fair, transparent, accountable, and private ai system

ICON

ESG Diagnostic
Calculating carbon emissions, minimizing environmental impact, and optimizing technological resources

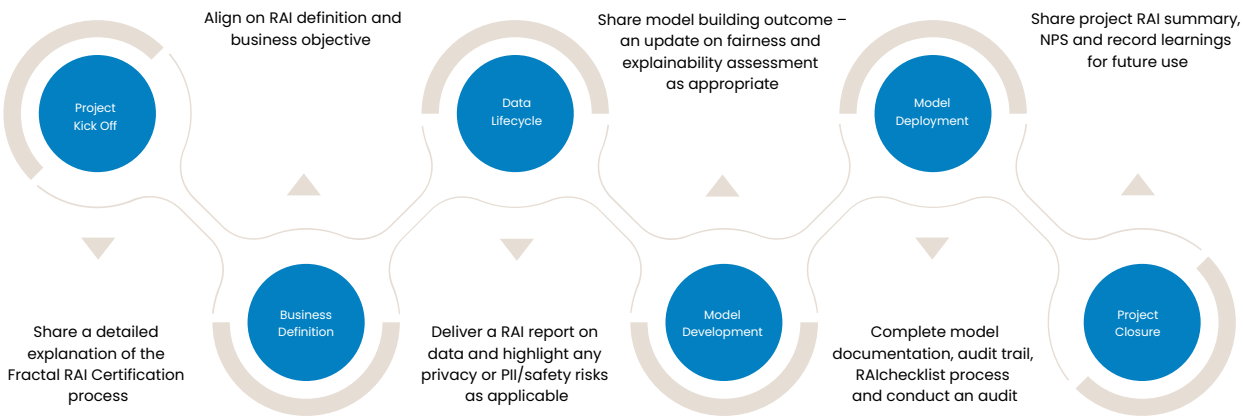
ICON

Synthetic data
Privacy by design and synthetic data for ethical, secure, and accountable practices

Fractal offers maturity assessments to evaluate clients' current standing regarding RAI practices and propose tailored solutions accordingly. From initial As-Is and To-Be assessments to in-depth diagnosis, we aim to foster collaboration and closely partner with organizations to incorporate RAI strategies into their systems.

To ensure RAI compliance, our certification process involves assessing projects using over 50 qualitative and quantitative criteria. Only after meeting these requirements is a project declared as RAI compliant.

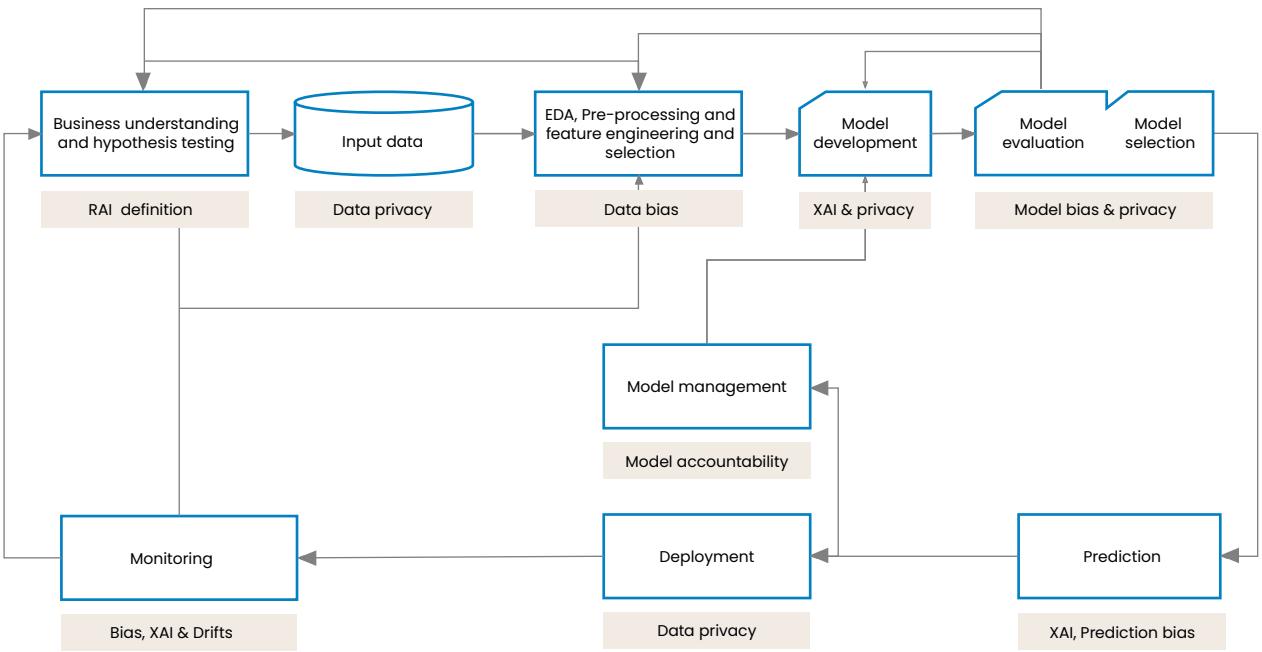
RAI Certification Process Flow



Fractal: Empowering better decisions

Remarkably, every typical AI life cycle stage presents opportunities and requirements for incorporating RAI modules. There is no stage where RAI modules can be disregarded, and it's crucial to embed RAI into the system right from the start rather than adding it as an afterthought once the AI development process has started.

Data science lifecycle with RAI



Applying our enablers, we have developed a comprehensive and reusable toolkit to integrate RAI into clients' data science life cycles seamlessly. Our codified frameworks, policies, guidelines, and cloud-agnostic codes can be applied across any data and data science process. By democratizing these resources through educational programs, we empower clients to effectively embed RAI principles and practices into their workflows and organizational practices.

Our toolkit includes the following components:

1. EMBEDDING RAI FRAMEWORK

Integrating the RAI framework and its underlying principles into AI systems to ensure responsible practices are integrated from the start.

2. GUIDELINES AND RECOMMENDATIONS:

Providing comprehensive guidelines and recommendations to guide clients in implementing RAI effectively.

3. MATURITY ASSESSMENT:

Conduct a thorough maturity assessment to evaluate clients' RAI adoption, score their status, and identify areas for improvement.

4. RAI TRAINING AND WORKSHOPS:

Creating RAI training modules and conducting workshops to promote awareness and evangelize RAI principles among stakeholders.

5. CODE MODULES:

Developing code modules specifically designed to address fairness, explainability, privacy, and drifts, supporting the implementation of RAI.

6. THOUGHT LEADERSHIP:

Demonstrating thought leadership in RAI through participation in conferences, workshops, and podcasts, sharing insights, and promoting best practices.

7. RAI-COMPLIANT USE CASES:

Building a repository of RAI-compliant use cases and certifying all projects to ensure adherence to RAI standards.

8. INTEGRATION WITH ORGANIZATIONAL DATA SCIENCE PRACTICES:

Integrating RAI into larger organizational data science practices and workflows, ensuring RAI becomes an integral part of the overall framework.

As the power and pitfalls of artificial intelligence emerge increasingly into the spotlight, the demand for Responsible AI (RAI) solutions is on the rise. Fractal's innovative Responsible AI 2.0 framework takes a forward-thinking approach by incorporating General Artificial Intelligence (GAI) to deliver meaningful benefits in risk management, underwriting, customer satisfaction, and, ultimately, the bottom line—Trust Fractal to provide game-changing RAI solutions that enable your business to thrive in the era of AI.

Authors

Ashna Taneja
Consultant,
Fractal Dimension

Sray Agarwal
Principal Consultant,
Fractal Dimension

About Fractal

Fractal is one of the most prominent providers of Artificial Intelligence to Fortune 500® companies. Fractal's vision is to power every human decision in the enterprise, and bring AI, engineering, and design to help the world's most admired companies.

Fractal's businesses include Crux Intelligence (AI driven business intelligence), Eugenie.ai (AI for sustainability), Asper.ai (AI for revenue growth management) and Senseforth.ai (conversational AI for sales and customer service). Fractal incubated Qure.ai, a leading player in healthcare AI for detecting Tuberculosis and Lung cancer.

Fractal currently has 4000+ employees across 16 global locations, including the United States, UK, Ukraine, India, Singapore, and Australia. Fractal has been recognized as 'Great Workplace' and 'India's Best Workplaces for Women' in the top 100 (large) category by The Great Place to Work® Institute; featured as a leader in Customer Analytics Service Providers Wave™ 2021, Computer Vision Consultancies Wave™ 2020 & Specialized Insights Service Providers Wave™ 2020 by Forrester Research Inc., a leader in Analytics & AI Services Specialists Peak Matrix 2022 by Everest Group and recognized as an 'Honorable Vendor' in 2022 Magic Quadrant™ for data & analytics by Gartner Inc.

For more information, visit fractal.ai

Corporate Headquarters
Suite 76J,
One World Trade Center, New York,
NY 10007

Get in touch