genpact

AI Responsible Artificial Intelligence Institute

Responsible AI

a strategic guide to address the elephant in the room

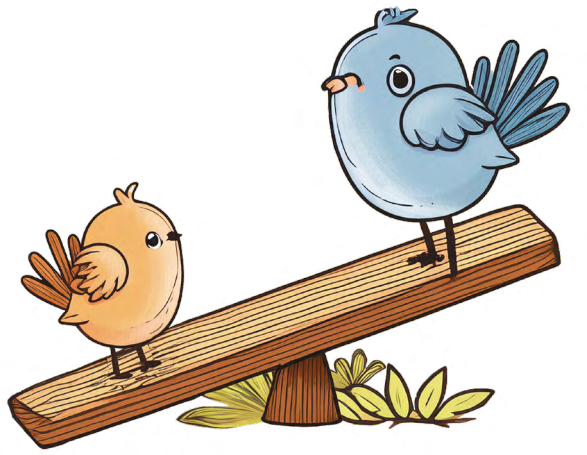

Joe, you are blocking the view!

TABLE OF CONTENTS

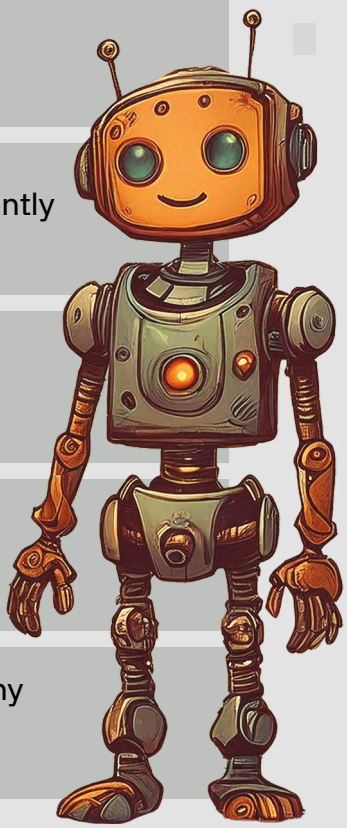
Why should one care about responsible AI?

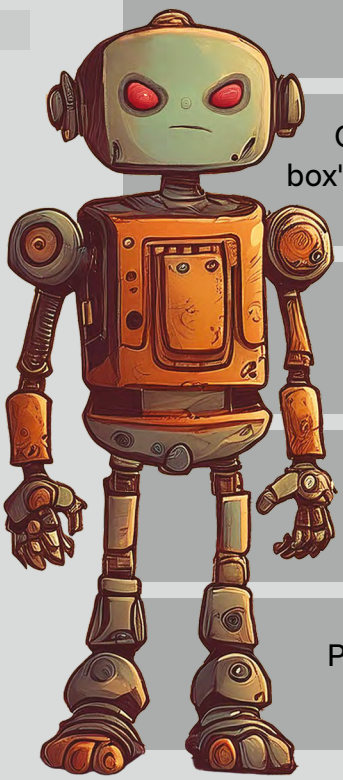
01

GOOD AI VS BAD AI



GOOD AI	BAD AI
Enhances human capabilities	Replaces humans without thought
Operates transparently and explicably	Operates as a "black box" with unexplainable decisions
Respects privacy and security	Plays fast and loose with data privacy
Promotes fairness and reduces bias	Perpetuates or amplifies existing biases
Aligns with company values and ethical standards	Prioritizes profit over people and ethics





Imagine you're at an airport, waiting for your flight. You're greeted by an AI-powered kiosk. This AI assistant swiftly processes your information and assigns you a window seat. You breeze through security. This is good AI at work.

Now imagine a scenario where the AI system is tasked with analyzing medical history and recommending treatment for your friend. However, this AI hasn't been properly trained on diverse patient data or rigorously tested for biases. It misses a crucial detail in your friend's medical history, leading to a potentially life-threatening misdiagnosis. This is AI doing badly.

Responsible AI as the name suggests, helps keep us in the lane, within the premises of responsibility and ethicality. It incentivizes good behavior in AI. Good and bad in the context of AI can be subject to interpretations. The goalposts can vary with industry, geography and even with use cases.

ONE OF THE KEY CHALLENGES IN AI SAFETY IS ENSURING THAT ADVANCED AI SYSTEMS ARE ALIGNED WITH HUMAN VALUES AND INTERESTS.

The difference between good and bad AI isn't just a matter of convenience versus catastrophe. It's about the fundamental approach to AI development and deployment. This handbook tries to carefully define safe AI systems and highlight often overlooked aspects of building AI solutions in an enterprise.

By now, we all have used or heard about Generative AI through ChatGPTs of the world. While few of us marveled at the potential of these GenAI applications, the others can't help but notice a fumbling and mumbling dumber counterpart (or assistant) to human intelligence. While the flourish or perish rhetoric in the context of AI is good for news headlines, building functional enterprise AI is rigorous, boring, and goes through lots of safety nets before it is exposed to the world. When these two worlds of sensationalism and objectivism collide, there is typically lot of confusion, which is atypical of the early stage hype cycle of any new technology or innovation.

AI, after decades of ups and downs has finally become a household name thanks to OpenAI's ChatGPT. The result is, AI is now more prone to malicious usage. Not to forget, the hallucinations. Sometimes, AI

goes a bit off script and starts seeing things that aren't there – we call this "hallucinating" which means a model has drifted away from the data it's been given or has added additional information not contained in it. There are times when hallucinations are beneficial, like when users want AI to create a science fiction story. But many organizations building AI copilots and chatbots need them to deliver reliable, grounded information in scenarios like medical summarization or investment advisory.

AN AI SYSTEM RUSHED TO MARKET WITHOUT PROPER SAFEGUARDS COULD BE A DISASTER WAITING TO HAPPEN.

Deepfake impersonations are increasingly prevalent in video calls and live streams, using digital footprints like facial data to create realistic personas that deceive older adults without suspicion. [1]

An AI system rushed to market without proper safeguards could be a disaster waiting to happen. If it's not trained on a diverse enough dataset, it might perform poorly for certain demographic groups. If it's not designed with appropriate human oversight, it could make critical errors that go unchecked. The consequences could range from harmful treatments to irreversible banking errors and more!

Therefore, Responsible AI isn't just a nice-to-have – it's a critical business imperative. It's the difference between an AI system that propels an organization forward and one that becomes a ticking time bomb of legal, ethical, and financial risks.

WITH AI, TRUST IS A MUST AND NOT JUST A NICE-TO-HAVE

2024

The European Artificial Intelligence Act (AI Act) enters into force. The Act aims to foster responsible artificial intelligence development and deployment in the EU. [9]

2023

The US government announced new actions to promote Responsible AI Innovation that protects Americans' rights and safety. [8]

2021

The EU proposes the Artificial Intelligence Act, the world's first comprehensive attempt to regulate AI systems based on their level of risk. [7]

2018

Google launches the AI Ethics Board to oversee the implementation of these principles and address ethical concerns. [6]

2016

Microsoft establishes the Aether Committee (AI and Ethics in Engineering and Research) to provide guidance on responsible AI development. [5]

2015

Elon Musk, and other prominent figures signed an open letter calling for research on the societal impacts of AI. [4]

2002

Philosopher Nick Bostrom published a seminal paper on existential risks from artificial intelligence, kickstarting serious academic discussion about AI safety. [3]

1985

Computer scientist Joseph Weizenbaum publishes "Computer Power and Human Reason," one of the first books to seriously examine the ethical implications of AI. [2]

EVOLUTION OF THE RESPONSIBILITY RHETORIC

Modern computing pioneer Alan Turing once asked, "Can machines think?" --- this set the stage for decades of philosophical and practical debates about AI's potential and pitfalls. In 1985, computer scientist Joseph Weizenbaum published "Computer Power and Human Reason," one of the first books to seriously examine the ethical implications of AI.

As AI systems became more powerful and pervasive, concerns about their impact grew. In 2002, philosopher Nick Bostrom published a seminal paper on existential risks from artificial intelligence, kickstarting serious academic discussion about AI safety.

2012 was a watershed year for AI, with the breakthrough of deep learning techniques that dramatically improved AI performance in tasks like image and speech recognition. This "AI spring" brought renewed attention to the potential and risks of AI.

In the same year, Google began publishing research papers on Responsible AI, covering topics like user perceptions, data protection, and adversarial testing.

This "AI spring" brought renewed attention to the potential and risks of AI. In 2015, Stephen Hawking, Elon Musk, and other prominent figures signed an open letter calling for research on the societal impacts of AI. The year 2016 saw the launch of the Partnership on AI, a coalition of tech companies, academics, and civil society organizations dedicated to the responsible development of AI. This marked a shift towards industry self-regulation and collaborative efforts to address AI ethics.

The following year, Microsoft established the Aether Committee (AI and Ethics in Engineering and Research) to provide guidance on responsible AI development. As the tech giants warmed up to the notion of ethical AI, in 2018, the European Union (EU) implemented the General Data Protection Regulation (GDPR), which included provisions

specifically addressing automated decision-making and profiling. This was one of the first major regulatory frameworks to grapple with AI-specific issues. Like Microsoft, Google launched the AI Ethics Board to oversee the implementation of these principles and address ethical concerns in the year 2018.

As a follow-up to their GDPR implementation, in 2021, the EU proposed the Artificial Intelligence Act, the world's first comprehensive attempt to regulate AI systems based on their level of risk. This marked a new phase in the Responsible AI journey, with governments taking a more active role in shaping AI development and use. But, little did they know what was to come!

In the fall of 2022, November, OpenAI released ChatGPT to the public that led to an explosive growth of generative AI, with models like GPT-3 and DALL-E capturing the public imagination. This brought Responsible AI concerns to the forefront of public discourse, with debates raging about AI's impact on jobs, creativity, and information integrity. As a response, in May 2023, the US government announced new actions to promote Responsible AI Innovation that protect Americans' rights and safety.

By 2024, companies like Amazon committed to continued collaboration with the White House, policymakers, technology organizations, and the AI community to advance the responsible and secure use of AI. On 1 August 2024, the European Artificial Intelligence Act (AI Act) entered into force. The Act aims to foster responsible artificial intelligence development and deployment in the EU.

Responsible AI has evolved from a niche concern to a mainstream business imperative.

The journey of AI responsibility started with early concerns about automation and job loss, moved through the ethical dilemmas highlighted by advanced machine learning, and now focuses on governance, transparency, and fairness. Each milestone reflects our growing understanding of AI's potential and the need for its ethical deployment.

Responsible AI has evolved from a niche concern to a mainstream business imperative. But numbers only tell part of the story. It's a story that's still being written, with each of us – as business leaders, technologists, and citizens – playing a crucial role in shaping the next chapter. As we look to the future, one thing is clear: Responsible AI isn't just a trend or a compliance checkbox. It's the key to unlocking AI's full potential while mitigating its risks.

Who watches the watchmen?

02

GATEKEEPERS OUTSIDE THE ENTERPRISE

> The EU Act, allows the EU to slap penalties as high as 7% of the global revenue of the companies in violation of the law. Companies utilizing LLMs, especially for high-risk applications will face stricter regulations. They'll need to demonstrate responsible data practices, ensure transparency about LLM use, and potentially explain how these models reach their outputs.

NIST AI 600-1: Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, was released on July 26, 2024, and serves as a companion resource to the broader NIST AI Risk Management Framework introduced in January 2023. [10]

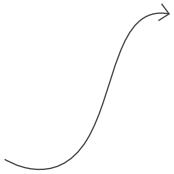
The EU AI Act was published in the EU Official Journal on July 12, 2024, and is the first comprehensive horizontal legal framework for the regulation of AI across the EU. The EU AI Act entered into force on August 1, 2024, with most provisions entering into effect from August 2, 2026

The Unites States' Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence from 2023 establishes a coordinated Federal Government-wide approach to governing AI.

The California AI Transparency Act (SB 942 requires covered providers to develop AI detection tools, offer users options for disclosure in AI-generated content, and revoke licenses for third-party licensees who modify AI systems to avoid disclosures. The bill is set to go into effect on January 1, 2026. [11]

Effective February 1, 2026, the Colorado AI Act mandates that developers and deployers of high-risk AI systems exercise reasonable care to prevent algorithmic discrimination, with compliance presumed if they follow specified practices such as transparency, risk assessment, and public disclosures. Developers must inform deployers and regulators of risks and maintain documentation, while deployers must implement risk management policies, allow consumer appeals, and disclose AI interactions. The act offers affirmative defenses for those adhering to recognized AI risk frameworks and exempts certain regulated financial and insurance entities. [12]

Today, with the passing of AI acts, AI-first companies have an obligation to conduct impact assessments to evaluate potential consequences and make sure that their AI development teams are informed about the Act's implementation and that any changes are vital. The EU AI Act, NIST RMF framework, and other similar initiatives are significant steps towards ensuring that AI is used for the benefit of society while mitigating its potential harms. The EU AI Act, for instance, categorizes AI systems based on their risk level and imposes varying regulatory requirements accordingly. This risk-based approach aims to strike a balance between innovation and safety. For instance, the EU Act, allows the EU to slap penalties as high as 7% of the global revenue of the companies in violation of the law. [13].



7%

of the global revenue of the companies in violation of the law

> Historically, 'red teaming' was a method to assess security weaknesses by simulating attacks. Today, 'red teaming' often refers to any kind of rigorous testing or probing of AI systems.

THE ONES IN-HOUSE

As enterprises race to integrate GenAI into their operations, they have an uphill task of navigating a complex landscape of governance and compliance to ensure responsible deployment. A key aspect of managing risk in GenAI adoption is the adaptation of existing governance structures. This approach minimizes disruption to decision-making processes while maintaining clarity in accountability. Central to effective risk management is the establishment of robust governance mechanisms. This includes forming cross-functional, responsible AI working groups comprising business and technology leaders, as well as experts in data, privacy, legal, and compliance domains. These groups serve as forums for collaboration, enabling proactive identification and mitigation of potential risks associated with GenAI deployment.

Additionally, organizations are encouraged to appoint dedicated AI governance officers to spearhead risk management efforts. These officers play a crucial role in coordinating governance activities, ensuring compliance with regulatory requirements, and fostering a culture of responsible AI within the organization. AI governance officers form teams to set governance frameworks and core risk frameworks and guide designers and engineers.

Internal compliance measures help ensure that high-quality decisions align with organizational and societal values. Achieving AI compliance and governance within an organization is like childproofing a house - one needs to anticipate problems before they happen and put safeguards in place.

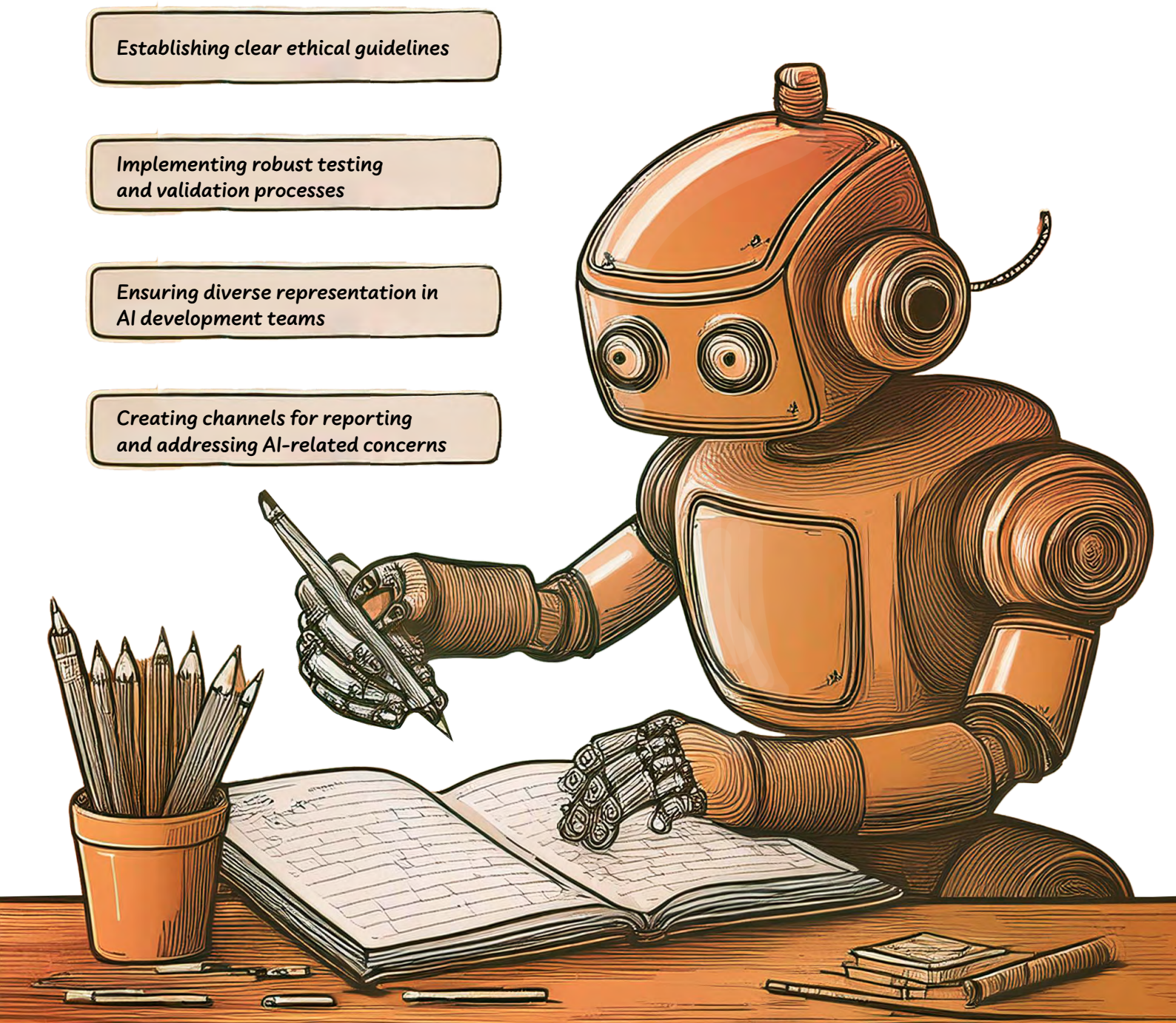
> “Responsible AI principles help marry commerciality with conscientiousness and build safe human-centric AI systems.”

Enterprises should revisit how governance is done across the board by establishing cross-functional working groups comprising AI governance and AI policy officers. They challenge risk assessment and coordinate "red team" tests with engineers, which play a vital role in risk identification. Red teams are people tasked with trying to break or misuse AI

systems to identify vulnerabilities. They're the professional troublemakers, always asking, "But what if...?"

That said, red teams alone won't solve all the edge cases. A team of even 1000 in a company like Google or Microsoft cannot anticipate what a billion customers would use their GenAI applications for. There is always a constant channel of human feedback that is captured to make a system better. Responsible AI principles help marry this commerciality with conscientiousness and build safe human-centric AI systems.

Some key strategies for internal compliance and governance include:



Why AI enterprises are better off with RAI?

03

RAI: RECIPE FOR SAFE AI

AI Governance goes beyond just utilizing data effectively. It's about building trust and ensuring responsible AI development. Imagine a self-driving car trained on inaccurate GPS data – the consequences could be fatal. Data quality standards ensure the data used for training generative AI models is accurate, complete, consistent, timely, and reliable. One way to achieve this is by implementing a unified data and AI governance process through Responsible AI frameworks.

Responsible AI framework offers 360-degree coverage of a typical AI lifecycle.

1 Industry-Specific Evaluation of Business Metrics

Imagine a tech company that launches a generative AI project without syncing it with their business goals. It's like serving champagne at a backyard barbecue—sure, it's fancy, but it doesn't quite fit. RAI framework marries business metrics with AI performance metrics, ensuring stakeholders can confidently implement solutions and measure success. This alignment is key to achieving tangible business outcomes.

2 Data Drift Mitigation

Data drift is like your favorite coffee shop changing its flavor overnight—nobody wants that kind of surprise. Think of an online retailer whose recommendation engine suddenly starts suggesting winter coats in summer. That's data drift in action. Mitigating this requires setting strict metrics for data quality and performance. To combat data drift, industry subject-matter experts set metrics for data quality, anonymization, and overall performance. This proactive approach ensures your AI systems stay relevant and accurate.

3 Reliability and Safety

Focusing on reliability means selecting and fine-tuning models for consistent outputs. Safety isn't expensive, it's priceless. In the AI world, reliable outputs are priceless. The RAI framework offers guidance on selecting and fine-tuning models helps enterprises achieve consistent and reliable outputs, avoiding the pitfalls of AI hallucinations.

4 Privacy and Security

Using privacy-by-default frameworks allows organizations to conduct due diligence and enhance transparency, protecting AI systems, applications, and users. With solid privacy and security measures, you can accelerate AI development confidently, knowing your data and systems are safeguarded.

5 Explainability

Think of explainability and traceability as the breadcrumbs that lead you back to the source—no mysteries, just clarity. Auditing mechanisms validate and monitor generative AI prompt engineering throughout the user journey, making outputs easy to explain and trace.

6 Fairness and Legal Compliance

RAI framework helps companies assess data trustworthiness and take steps to reduce bias using internal standards and new government guidelines, such as the EU's AI Act. Fair play and legal compliance aren't optional—they're essential for building trustworthy AI systems.

7 Autonomy and Accountability

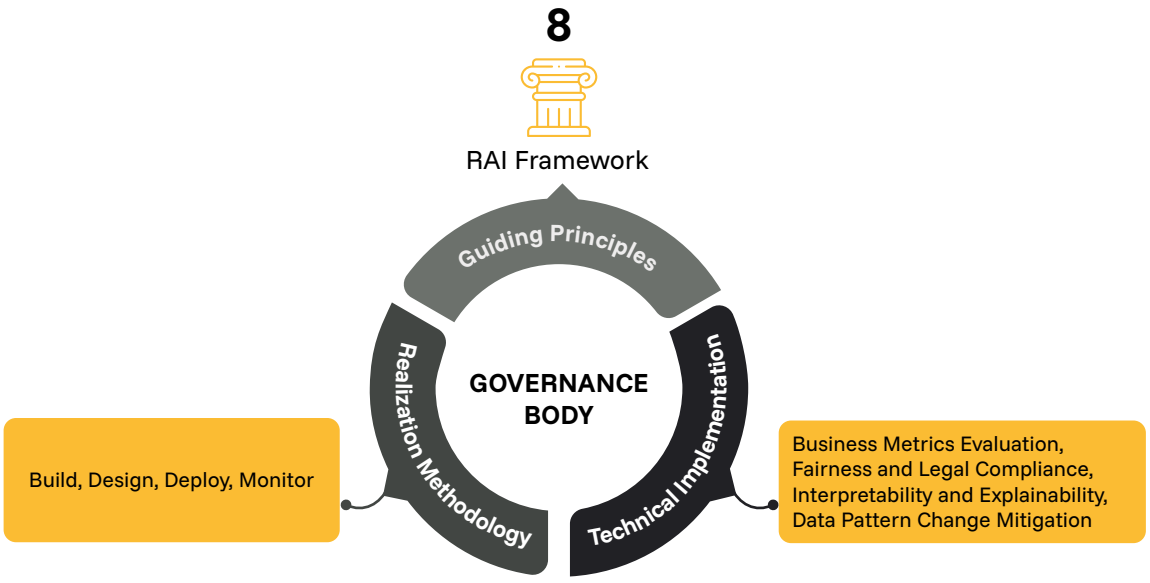
Autonomy refers to the definition of the level of control over outputs and decisions made by the AI System. Accountability refers to the ownership of the data, model and AI System and associated redressal policies.

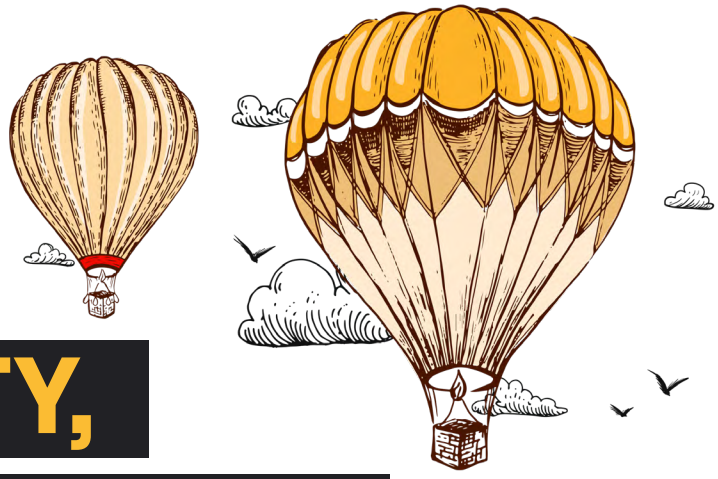
8 Traceability

Traceability is the ability of the system to track the output to the inputs and associated processing. Clarity builds trust. Ensure every AI decision is transparent and traceable, from start to finish.

This process establishes clear ownership, access controls, and auditing mechanisms for all data and AI assets within the organization. Choosing the right governance model, whether centralized or distributed, depends on the specific needs of the organization, but having a system in place is paramount. Security is another crucial aspect of data governance. Centralized access control for all data and AI assets minimizes the risk of unauthorized access or manipulation. These features are vital for maintaining a robust security posture and mitigating potential risks.

Responsible AI Framework





RAI IS AN OPPORTUNITY, NOT A ROADBLOCK

For more than a third of this handbook, we have been preparing and warming ourselves up to the world of responsible and trustworthy AI. We have discussed AI safety through the lens of governments and the general public, through their regulations and their naivety. Some might even argue, especially the LLM-makers, against the regulatory practices and view Responsible AI as a hindrance to innovation. We cannot blame them. Such is the pace at which GenAI models and use cases are being improved while the regulations play a catch-up.

For AI service providers who have to participate with stakeholders from both sides of the fence, Responsible AI is an opportunity. It's a bridge--- more of a tightrope--between two distant islands. Companies like Genpact have had decades of experience of building solutions for top healthcare, banking, retail and manufacturing customers. The subject matter expertise amassed over the years through experiments and conversations appropriately prepared us for the balancing act--- Responsible AI.

AI clientele want AI solutions to embody accountability, transparency, privacy, safety, security, and resilience. However, they face significant challenges in adopting new frameworks. Rather than starting from scratch, they're looking for AI governance solutions that can seamlessly integrate with their existing risk management infrastructure.

To meet these needs, AI service providers must establish governance mechanisms rooted in these key principles, while also educating customers on the opportunities and risks of AI. This involves developing practical implementation toolkits, building robust AI audit mechanisms, and leveraging in-house RAI experts to support customers' risk management teams. Success in this space hinges on offering comprehensive, integrated RAI solutions that align with customers' current processes and directly address their specific concerns. By providing not just cutting-edge technology, but also the expertise and support necessary to navigate the complexities of AI governance, service providers can help companies build trust, promote compliance, and drive responsible innovation. This positions them as invaluable partners in the journey towards sustainable and ethical AI adoption.

10 MANTRAS

that can Save You From Sounding Clueless in Client Meetings

How Not to Talk About RAI

- **Don't propose entirely new frameworks:** "You'll need to implement a completely new system."
- **Don't overwhelm the customer:** “Our multi-modal federated learning approach leverages homomorphic encryption for privacy-preserving neural network training.”
- **Don't make vague promises:** Avoid statements like, "Our AI is 100% safe and foolproof."
- **Don't offer piecemeal solutions:** "We can help with just one aspect of AI safety."
- **Don't just scratch the surface:** "Our whitepapers and webinars provide theoretical insights into AI governance."
- **Don't ignore the need for continuous monitoring:** “Once you implement our RAI solution you don’t have look back.”
- **Don't suggest customers need to become AI experts:** "You'll need to develop deep AI expertise internally."
- **Don't focus only on risks or only on benefits:** Avoid one-sided presentations of AI's impact.
- **Don't offer one-size-fits-all solutions:** "This is the standard RAI package that everyone uses."
- **Don't downplay the importance of RAI:** "RAI is just a nice-to-have feature."

How to Talk About RAI

- **Emphasize integration with existing systems:** "Our RAI solutions seamlessly integrate with your current risk management setup."
- **Focus on customer pain points:** "We understand the challenges you're facing with AI adoption, and our solutions address them directly."
- **Highlight key principles:** "Our RAI solutions are built on accountability, transparency, privacy, safety, security, and resilience."
- **Offer holistic solutions:** "We provide comprehensive AI governance solutions that cover all aspects of responsible AI."
- **Offer more than thought leadership:** "We help you navigate the opportunities and risks of AI, offering education and practical tools for your teams."
- **Discuss audit mechanisms:** "Our robust AI audit mechanisms facilitate ongoing compliance and improvement catering to the EU AI Act, NIST etc"
- **Mention expert support:** "Our in-house RAI experts are available to empower your risk management teams."
- **Address both opportunities and risks:** "We'll help you navigate both the prosperity and perils of AI."
- **Emphasize customization:** "Our RAI solutions can be tailored to your specific industry and organizational needs."
- **Highlight the increasing importance of AI safety:** "With the rise of large language models, prioritizing AI safety has never been more crucial."

Don’t pass the buck!

04

C-SUITE OBLIGATION AND ACCOUNTABILITY

Responsible AI isn't just an IT issue or a legal concern – it's a fundamental business imperative that touches every aspect of the organization. Leaders have a unique obligation to champion and drive Responsible AI initiatives.

First and foremost, Responsible AI is about risk management. In today's digital age, an AI misstep can cost dearly. We keep hearing stories of companies that have lost millions in market value after their AI system was found to be making biased decisions. By prioritizing Responsible AI, companies are not just protecting the bottom line – they're safeguarding the company's reputation and future.

According to a recent survey, 80% of business leaders see explainability, ethics, bias, or trust as a major concern on the road to generative AI adoption.. Some organizations may cut corners to move ahead quickly, but most are committed to responsible action. In fact, the survey states that 72% of executives say they'll step back from generative AI initiatives if they think the benefits could come at an ethical cost. In this case, less is more. These same organizations are 27% more likely to outperform on revenue growth than others. [14]

The role of a C-suite executive in driving Responsible AI is multifaceted. It starts with setting the tone from the top. Their words and actions signal to the entire organization that Responsible AI is a priority. This means going beyond lip service – it means allocating resources, setting clear policies, and holding people accountable for ethical AI practices.

C-suite role goes beyond lip service – it means allocating resources, setting clear policies, and holding people accountable for ethical AI practices.

This might involve establishing an AI ethics board, implementing regular AI audits, or creating channels for employees and customers to raise concerns about AI systems. Culture eats strategy for breakfast – all the AI policies in the world won't help if the organization doesn't have a culture that values responsible innovation.

Another crucial aspect of an organization's Responsible AI obligation is ensuring proper governance structures are in place. This might involve creating new roles (think: Chief AI Ethics Officer) or new processes for AI development and deployment. It's about creating checks and balances to ensure that AI systems are aligned with the organization's values and ethical principles.

Education is also key. As a leader, one needs to ensure that the entire organization – from the boardroom to the frontlines – understands the basics of AI and its ethical implications. This isn't about turning everyone into AI experts, but about creating a shared language and understanding around Responsible AI. Studies show that organizations that invest in AI literacy see five times the return on their AI investments compared to those that don't.

> A dynamic AI governance framework and collaboration across sectors are vital for developing effective, universally accepted Responsible AI tools.

The C-suite role is to ask the tough questions. When a new AI system is proposed, one should be asking:

It's also crucial to recognize that Responsible AI isn't a one-and-done effort. The AI landscape is evolving at breakneck speed, with new technologies, use cases, and ethical dilemmas emerging all the time. This might mean regularly reviewing and updating the AI ethics policies, or partnering with academic institutions or think tanks to stay ahead of the curve.

Finally, remember that Responsible AI isn't just an internal matter – it's about the entire ecosystem. As a leader, one has an obligation to engage with stakeholders, from customers to regulators to the broader public, about the AI practices. This transparency not only builds trust but can also give one valuable insight to improve Responsible AI efforts.

Here are a few C-suite best practices for implementing RAI effectively

Build Awareness of Responsible AI

Your responsible AI framework must be a living, breathing part of your organization. This means creating a communication plan that not only spreads awareness but also enforces these principles across every department. It's about making responsible AI as integral to your enterprise as your morning coffee. Don't let your responsible AI framework be the best-kept secret in your organization. Share it wide, make it part of your culture, and watch as it transforms your enterprise.

Prepare in Advance

"By failing to prepare, you are preparing to fail," said Benjamin Franklin. Before rolling out your gen-AI solution, map out exactly which processes it will touch. This includes setting up actions to address any legal, security, or ethical concerns that might pop up. It's like having a fire extinguisher on hand before you start a campfire. Preparation isn't just a good idea; it's essential. Know your AI's impact, anticipate the risks, and plan your mitigation strategies accordingly.

Rally Around Common Benefits

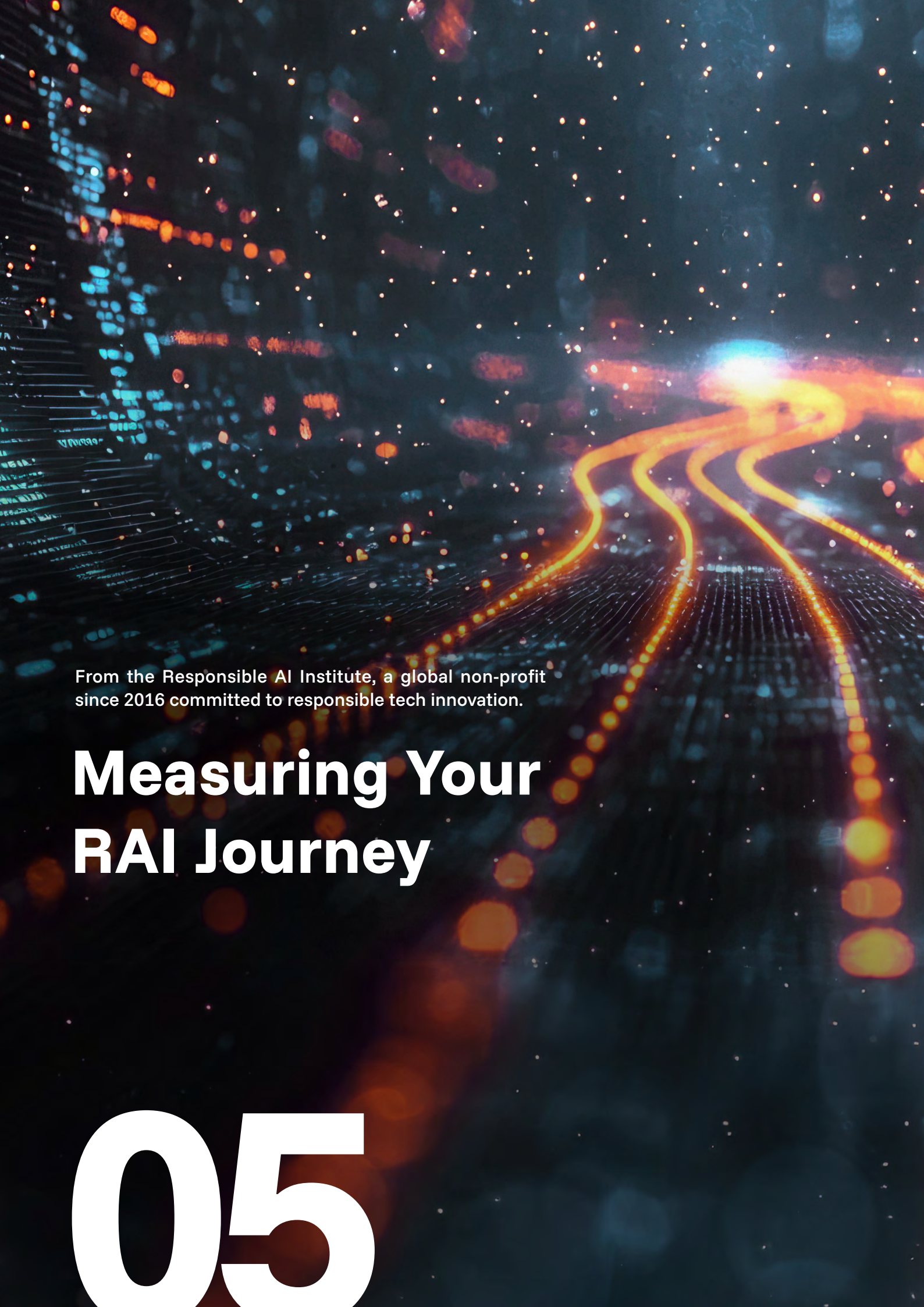
"Transparency is the currency of trust in the digital age," says Daniel Newman. And for generative AI, this couldn't be more true. Transparency inside and outside your enterprise is key when adopting generative AI. Clearly communicate the benefits to all stakeholders and be authentic. Whether it's improved efficiency, better customer service, or enhanced decision-making, make sure everyone's on the same page. To get everyone on board with generative AI, highlight the common benefits and communicate them transparently. It's a dance everyone will want to join.

Prioritize Explainability

To gain trust, your generative AI tools need to be accessible and explainable. Use the right resources, libraries, and frameworks to show how AI arrives at its decisions. Make your AI's decision-making process as transparent as possible. Explainability is the key to earning and maintaining stakeholder trust.

Embed Reliability Metrics

Formulate confidence scores for your generative AI outputs and pass them through human evaluation. This human-in-the-loop approach allows expertise to fine-tune algorithms and improve accuracy, fostering trust in your AI solutions. Reliability metrics are your AI's trust badges. Use them wisely, and ensure human oversight to keep your AI on the right track.



From the Responsible AI Institute, a global non-profit since 2016 committed to responsible tech innovation.

Measuring Your RAI Journey

05

MEASURING YOUR RAI JOURNEY: WHERE DOES YOUR ORGANIZATION STAND?

Remember Joe blocking everyone's view in that boardroom? A lot of organizations are in the same boat when it comes to their AI efforts. Some are still trying to figure out what the elephant even is. Others have learned to work around it. A few have actually invited it onto the team. So—where does your organization stand?

The Responsible AI Institute (RAI Institute) has been in the weeds of this question for almost a decade. After studying how organizations build (and stumble through) responsible AI efforts, they noticed a pattern. Despite all the different industries, sizes, and tech stacks, most companies move through five familiar stages. It's not always linear. And it's rarely neat. But it's a useful way to make sense of the mess.

Knowing your RAI stage helps you identify gaps before auditors or regulators do, build the case for change across departments, and prioritize resources where they'll have real impact

The Five Stages: From Chaos to Clarity

1

Aware – "Wait, There's an Elephant?"

This is where most people begin. A few folks are raising concerns about AI bias or privacy, but it's scattered and unofficial. The organization knows there's something to pay attention to—just not what to do about it yet.

2

Active – "Let's Make a Plan (Sort Of)"

Things start to take shape. A policy here, a fairness tool there. People are doing the work, but mostly in silos. Legal might not know what engineering is doing, and nobody's sure who's actually in charge. But hey, at least the elephant has a name now.

3

Operational – "We've Got This (Mostly)"

This is where things start to click. Teams talk to each other. There are shared goals, repeatable processes, and someone who actually owns the work. Responsible AI is no longer a side quest—it's becoming part of how things get done.

4

Systemic – "RAI? It's in Our DNA"

At this point, RAI is baked into decision-making. It's part of onboarding, procurement, design reviews, and business strategy. People don't need a reminder to "think about the ethics"—they just do.

5

Transformative – "We're Helping Others Tame Elephants"

Some organizations go a step further. They not only manage their own AI risks—they help others do the same. They're contributing to standards, sharing lessons, and showing what good can look like at scale.

Plot Twist: Maturity Isn't About Size or Budget

A two-year-old startup can be further along than a 50-year-old multinational. It's less about headcount and more about mindset. The organizations that make real progress are the ones willing to ask hard questions, invest in long-term thinking, and get comfortable with the uncomfortable.

Some companies move through the stages in a year. Others get stuck in Stage 2 for what feels like forever. That's okay. What matters most is knowing where you are—and being honest about what needs to change.

> "Eight years ago, I founded the Responsible AI Institute with a vision to ensure AI is deployed responsibly. Today, as adoption accelerates, organizations are grappling with the challenge of balancing innovation with accountability. The most forward-looking leaders now understand: Responsible AI is not a roadblock to innovation—it's the infrastructure for it. When we confront the elephant in the room with mature, systematic governance, we don't just reduce risk—we enable trust, scale, and sustainable impact."
>
> **Manoj Saxena, Founder and Executive Chairman,**
> **Responsible AI Institute**

So, What Now?

This model isn't a scoreboard. You don't win by reaching Stage 5. The goal is to use it as a mirror—to figure out what's working, what's missing, and what comes next.

Most organizations today hover somewhere between Active and Operational. That's not a problem—it's a starting point. The key is to move forward with purpose, not panic.

GOING FORWARD

Initiatives such as OpenAI's GPT-4o System Card indicate big tech's commitment to responsible AI development. The System Card provides a comprehensive view of the model's development process, safety considerations, and evaluation methods, aligning well with emerging regulatory frameworks like the EU's AI Act. The phased development approach of progressively introducing complex capabilities while implementing safety measures, serves as a valuable industry model. It demonstrates how we can push AI capabilities while prioritizing safety and ethics at each step. The emphasis on extensive testing throughout development, including red teaming exercises to identify risks and develop mitigations, is a practice worth adopting industry-wide. While the implementation of output classifiers and moderation systems is a step forward, there's still work needed in developing standardized methods for marking AI-generated content. That said, the evaluation approach is beginning to evolve as well. By addressing issues such as underspecified problem statements and overly specific unit tests, we can create more accurate assessments of AI performance. Removing impossible or ambiguous tasks gives us a clearer picture of true AI capabilities, allowing for more informed decisions about model deployment and risk mitigation. The AI industry is making significant strides in preparedness and evaluation. As service providers, it is key to invest in understanding the evolving benchmarks. This includes considering potential external enhancements and ecosystem-wide progress when assessing risks.

At the end of the day, a well-defined responsible AI operating model should map out interactions among various personas throughout the AI life cycle, tailored to each organization's capabilities.


GenAI tools
will become
as essential
to productivity
as coffee or
calculators

Glossary

RAI AND THE QUEST FOR PERFECT PIZZA

A

ACCOUNTABILITY: WHO GETS THE BLAME FOR THE BURNT CRUST?

Just like someone takes responsibility for ensuring the pizza is cooked perfectly, someone (or some team) should be accountable for the actions and decisions of an AI system.

BIAS IN AI: PEPPERONI OVERLOAD

B

Imagine your favorite pizza outlet, Pizza Perfect, adding pepperoni even if you prefer veggies last time. Bias in training data can lead to uninspired (and potentially undesirable) AI suggestions.

AI ALIGNMENT

Ensuring AI systems behave in ways that align with human values and intentions. Making sure the pizza chef (AI) creates pizzas exactly as the customers (humans) want, without adding unexpected toppings or using cooking methods we didn't approve.

D

DATA GOVERNANCE: FRESH TOPPINGS, FRESH DATA

Data used to train AI systems needs careful management, just like ensuring fresh, high-quality ingredients for a good pizza. Data governance ensures the data is clean, accurate, and protects user privacy.

E

EXPLAINABILITY & INTERPRETABILITY: WHY THE SURPRISE PINEAPPLE?

Understanding how AI arrives at its recommendations is crucial. Imagine seeing "pineapple" on your pizza even though you never ordered it! Explainability allows you to see the reasoning behind AI suggestions, just like knowing the ingredients.

H

HUMAN-IN-THE-LOOP (HIL): THE CHEF'S OVERSIGHT

Humans should be involved in developing, deploying, and monitoring AI systems. Think of a skilled pizza chef overseeing Pizza Perfect, ensuring it recommends perfect pizzas, not just cheese overload.

I

INCLUSIVENESS: BEYOND CHEESE AND PEPPERONI

Designing AI systems that consider diverse needs and perspectives. Imagine PizzaPerfect offering a wider range of toppings and options that cater to different dietary restrictions – not just the usual cheese and pepperoni combo.

F

FAIRNESS: MAKING SURE THOSE WHO DESERVE GET A SLICE!

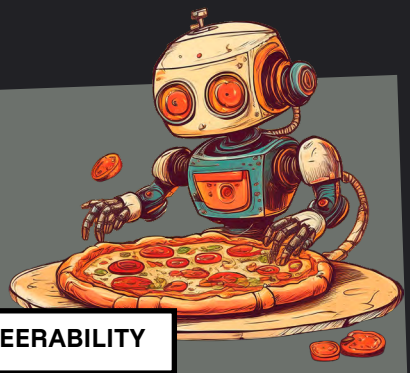
AI systems should be fair and inclusive, catering to diverse preferences. A responsible pizzaiolo offers options for everyone – vegetarians, meat lovers, gluten-free folks – so everyone gets a slice they enjoy!

GUARDRAILS

Restrictions or safety measures put in place to prevent AI systems from taking harmful actions. Pizza analogy: Installing a maximum temperature limit on the pizza oven to prevent it from ever getting hot enough to start a fire.

G

M



MODEL STEERABILITY

The ability to guide or control an AI model's outputs or behaviors in desired directions. Being able to easily adjust the pizza-making robot to create different styles of pizza (New York, Chicago, Neapolitan) without having to reprogram it completely.

P

PRIVACY: KEEPING YOUR FAVORITE TOPPINGS SECRET

Protecting user data from unauthorized access or use is paramount. Just like keeping your preferred pizza toppings private, user data should be handled with care in AI development

S

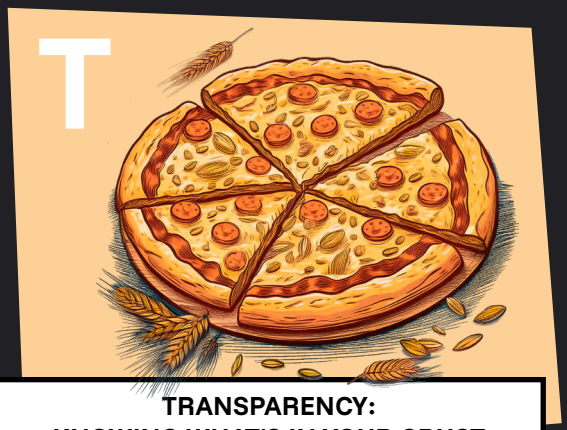
SECURITY: AVOIDING BURNT PIZZAS

Protecting AI systems from unauthorized access, use, disruption, or destruction is crucial. Imagine your pizza getting stolen on the way to your house! Secure AI systems prevent similar harm.

R

RELIABILITY AND SAFETY: A CONSISTENT, DELICIOUS PIZZA

Building AI systems that are dependable and risk averse is just like wanting a pizza that tastes good every time, we want reliable AI systems that function consistently and safely.



T

TRANSPARENCY: KNOWING WHAT'S IN YOUR CRUST

Transparency is about understanding how AI arrives at its conclusions. Just like knowing the ingredients on your pizza crust (whole wheat, white, etc.), users should be able to see the reasoning behind AI recommendations.

RED TEAMING- FINDING THE IMPOSSIBLE PIZZA

The practice of stress-testing AI systems by intentionally trying to make them fail or produce undesired outcomes. This is akin to hiring a group of kids to try every possible way to trick the automated pizza ordering system into making ridiculous or impossible pizzas.


Looks like Joe can finally see the road ahead!

Sreekanth Menon,
Global AI Practice Leader,
Genpact

It's tempting to treat responsibility as a compliance exercise. But in practice, responsible AI is infrastructure. It is embedded in how we log model behavior, control access, detect emergent drift, and embed escalation paths into automated systems. It is not static. It must evolve with the model landscape. The enterprise-wide rollout of RAI frameworks is supported through internal training programs, AI literacy efforts, and published design guidelines. This strategic guidebook is a good starter for AI teams looking to build RAI solutions.

Manoj Saxena,
Founder and Executive Chairman,
Responsible AI Institute

Eight years ago, I founded the Responsible AI Institute with a vision to ensure AI is deployed responsibly. Today, as adoption accelerates, organizations are grappling with the challenge of balancing innovation with accountability. The most forward-looking leaders now understand: Responsible AI is not a roadblock to innovation—it's the infrastructure for it. When we confront the elephant in the room with mature, systematic governance, we don't just reduce risk—we enable trust, scale, and sustainable impact.

REFERENCES

[1] Hear Us, then Protect Us: Navigating Deepfake Scams Zhai et al

[2] Reactions to Weizenbaum's Book

[3] Existential Risks Analyzing Human Extinction Scenarios and Related Hazards

[4] Open letter on artificial intelligence (2015)

[5] Advancing AI trustworthiness: Updates on responsible AI research

[6] Our commitment to advancing bold and responsible AI, together

[7] EU AI Act: first regulation on artificial intelligence

[8] Biden-Harris Administration Announces New Actions to Promote Responsible AI

[9] AI Act enters into force

[10] AI Risk Management Framework: NIST

[11] SB 942: California AI Transparency Act

[12] Consumer Protections for Artificial Intelligence

[13] European Artificial Intelligence Act comes into force

[14] The CEO’s guide to generative AI

DISCLAIMER

This Responsible AI Handbook (the "Handbook") is published by Genpact for general educational and informational purposes only. The AI landscape is rapidly evolving, and this Handbook reflects information available as of its publication date which may not reflect the most current developments and may be changed at any time. The content within this Handbook contains general industry knowledge and third-party sources and Genpact does not independently verify, validate or audit such data. Images and illustrations used in the Handbook were created using AI.

ALL INFORMATION IN THIS HANDBOOK, ANY RESULTS, INFORMATION, OR ANALYSIS IS PROVIDED "AS IS" WITHOUT ANY WARRANTY EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OR MERCHANTABILITY, FITNESS FOR PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT.

This Handbook is not intended to provide, and should not be construed as providing any legal, regulatory, compliance or other professional consultation or advice of any kind. This Handbook should not be used as a substitute for legal, regulatory, or business advice, nor a substitute for any detailed research or exercise of professional judgement.

Contact Information

AI ML SERVICES

aiml.services@genpact.com