What is responsible AI?

Explore IBM® watsonx.governance →

Published: 6 February 2024
Contributor: Cole Stryker

What is responsible AI?

Responsible artificial intelligence (AI) is a set of principles that help guide the design, development, deployment and use of AI—building trust in AI solutions that have the potential to empower organizations and their stakeholders. Responsible AI involves the consideration of a broader societal impact of AI systems and the measures required to align these technologies with stakeholder values, legal standards and ethical principles. Responsible AI aims to embed such ethical principles into AI applications and workflows to mitigate risks and negative outcomes associated with the use of AI, while maximizing positive outcomes.

This article aims to provide a general view of responsible AI. To learn more about IBM's specific point of view, see our AI ethics page.

The widespread adoption of machine learning in the 2010s, fueled by advances in big data and computing power, brought new ethical challenges, like bias, transparency and the use of personal data. AI ethics emerged as a distinct discipline during this period as tech companies and AI research institutions sought to proactively manage their AI efforts responsibly.

According to Accenture research: “Only 35% of global consumers trust how AI technology is being implemented by organizations. And 77% think organizations must be held accountable for their misuse of AI.”[1] In this atmosphere, AI developers are encouraged to guide their efforts with a strong and consistent ethical AI framework.

This applies particularly to the new types of generative AI that are now being rapidly adopted by enterprises. Responsible AI principles can help adopters harness the full potential of these tools, while minimizing unwanted outcomes.

AI must be trustworthy, and for stakeholders to trust AI, it must be transparent. Technology companies must be clear about who trains their AI systems, what data was used in that training, and, most importantly, what went into their algorithm’s recommendations. If we are to use AI to help make important decisions, it must be explainable.

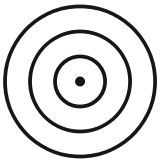
Related content

The Pillars of Trust

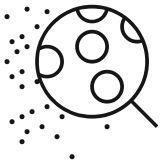
IBM has developed a framework to make these principles clear. Let's look at the properties that make up the “Pillars of Trust.” Taken together, these properties answer the question, “What would it take to trust the output of an AI model?” Trusted AI is a strategic and ethical imperative at IBM, but these pillars can be used by any enterprise to guide their efforts in AI.

Explainability

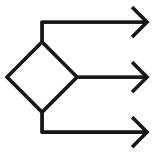
Machine learning models such as deep neural networks are achieving impressive accuracy on various tasks. But explainability and interpretability are ever more essential for the development of trustworthy AI. Three principles comprise IBM’s approach to explainability.



Prediction accuracy
Accuracy is a key component of how successful the use of AI is in everyday operation. By running simulations and comparing AI output to the results in the training data set, the prediction accuracy can be determined. The most popular technique used for this is Local Interpretable Model-Agnostic Explanations (LIME), which explain the prediction of classifiers by the machine learning algorithm.



Traceability
Traceability is a property of AI that signifies whether it allows users to track its predictions and processes. It involves the documentation of data and how it is processed by models. Traceability is another key technique for achieving explainability, and is accomplished, for example, by limiting the way decisions can be made and setting up a narrower scope for machine learning rules and features.



Decision understanding
This is the human factor. Practitioners need to be able to understand how and why AI derives conclusions. This is accomplished through continuous education.

Fairness

Machine learning models are increasingly used to inform high stakes decision-making that relates to people. Although machine learning, by its very nature, is a form of statistical discrimination, the discrimination becomes objectionable when it places privileged groups at systematic advantage and certain unprivileged groups at systematic disadvantage, potentially causing varied harms. Biases in training data, due to either prejudice in labels or under-/over-sampling, yields models with unwanted bias.

– **Diverse and representative data**

Ensure that the training data used to build AI models is diverse and representative of the population it is meant to serve. Include data inputs from various demographic groups to avoid underrepresentation or bias. Regularly check and assess training data for biases. Use tools and methods to identify and correct biases in the dataset before training the model.

– **Bias-aware algorithms**

Incorporate fairness metrics into the development process to assess how different subgroups are affected by the model's predictions. Monitor and minimize disparities in outcomes across various demographic groups. Apply constraints in the algorithm to ensure that the model adheres to predefined fairness criteria during training and deployment.

– **Bias mitigation techniques**

Apply techniques like re-sampling, re-weighting and adversarial training to mitigate biases in the model's predictions.

– **Diverse development teams**

Assemble interdisciplinary and diverse teams involved in AI development. Diverse teams can bring different perspectives to the table, helping to identify and rectify biases that might be overlooked by homogeneous teams.

– **Ethical AI review boards**

Establish review boards or committees to evaluate the potential biases and ethical implications of AI projects. These boards can provide guidance on ethical considerations throughout the development lifecycle.

Robustness

Robust AI effectively handles exceptional conditions, such as abnormalities in input or malicious attacks, without causing unintentional harm. It is also built to withstand intentional and unintentional interference by protecting against exposed vulnerabilities. Our increased reliance on these models and the value they represent as an accumulation of confidential and proprietary knowledge, are at increasing risk for attack. These models pose unique security risks that must be accounted for and mitigated.

Transparency

Users must be able to see how the service works, evaluate its functionality, and comprehend its strengths and limitations. Increased transparency provides information for AI consumers to better understand how the AI model or service was created. This helps a user of the model to determine whether it is appropriate for a given use case, or to evaluate how an AI produced inaccurate or biased conclusions.

Privacy

Many regulatory frameworks, including GDPR, mandate that organizations abide by certain privacy principles when processing personal information. A malicious third party with access to a trained ML model, even without access to the training data itself, can still reveal sensitive personal information about the people whose data was used to train the model. It is crucial to be able to protect AI models that may contain personal information, and control what data goes into the model in the first place.

Implementing responsible AI practices

Implementing responsible AI practices at the enterprise level involves a holistic, end-to-end approach that addresses various stages of AI development and deployment.

Define responsible AI principles

Develop a set of responsible AI principles that align with the values and goals of the enterprise. Consider the key aspects described above in the “Pillars of Trust.” Such principles can be developed and maintained by a dedicated cross-functional AI ethics team with representation from diverse departments, including AI specialists, ethicists, legal experts and business leaders.

Educate and raise awareness

Conduct training programs to educate employees, stakeholders and decision-makers about responsible AI practices. This includes understanding potential biases, ethical considerations and the importance of incorporating responsible AI into business operations.

Integrate ethics across the AI development lifecycle

Embed responsible AI practices across the AI development pipeline, from data collection and model training to deployment and ongoing monitoring. Employ techniques to address and mitigate biases in AI systems. Regularly assess models for fairness, especially regarding sensitive attributes such as race, gender or socioeconomic status. Prioritize transparency by making AI systems explainable. Provide clear documentation about data sources, algorithms, and decision processes. Users and stakeholders should be able to understand how AI systems make decisions.

Protect user privacy

Establish strong data and AI governance practices and safeguards to protect end user privacy and sensitive data. Clearly communicate data usage policies, obtain informed consent and comply with data protection regulations.

Facilitate human oversight

Integrate mechanisms for human oversight in critical decision-making processes. Define clear lines of accountability to ensure responsible parties are identified and can be held responsible for the outcomes of AI systems. Establish ongoing monitoring of AI systems to identify and address ethical concerns, biases or issues that may arise over time. Regularly audit AI models to assess compliance with ethical guidelines.

Encourage external collaboration

Foster collaboration with external organizations, research institutions, and open-source groups working on responsible AI. Stay informed about the latest

developments in responsible AI practices and initiatives and contribute to industry-wide efforts.

Related solutions

AI Ethics at IBM

IBM's multidisciplinary, multidimensional approach to trustworthy AI.

Explore AI Ethics at IBM →

IBM watsonx.governance

Build responsible, transparent and explainable AI workflows.

Explore IBM watsonx.governance →

AI governance services

IBM Consulting helps you weave responsible AI governance into the fabric of your business.

Explore IBM Consulting: AI Governance →

Resources

Blog

A look into IBM's AI ethics governance framework

IBM has publicly defined its multidisciplinary, multidimensional approach to AI ethics, built upon principles for trust and transparency.

Article

IBM's Principles for Trust and Transparency

For more than a century, IBM has earned the trust of our clients by responsibly managing their most valuable data, and we have worked to earn the trust of society by ushering powerful new technologies into the world responsibly and with clear purpose.

Article

Explainable AI

Explainable artificial intelligence (XAI) is a set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms.

Take the next step

Accelerate responsible, transparent and explainable AI workflows across the lifecycle for both generative and machine learning models. Direct, manage, and monitor your organization's AI activities to better manage growing AI regulations and detect and mitigate risk.

Explore watsonx.governance →

Book a live demo →

Footnotes

[1] Technology Vision 2022 (link resides outside ibm.com), Accenture, 2022.