Responsible AI - Recommendations

Published on September 2, 2019

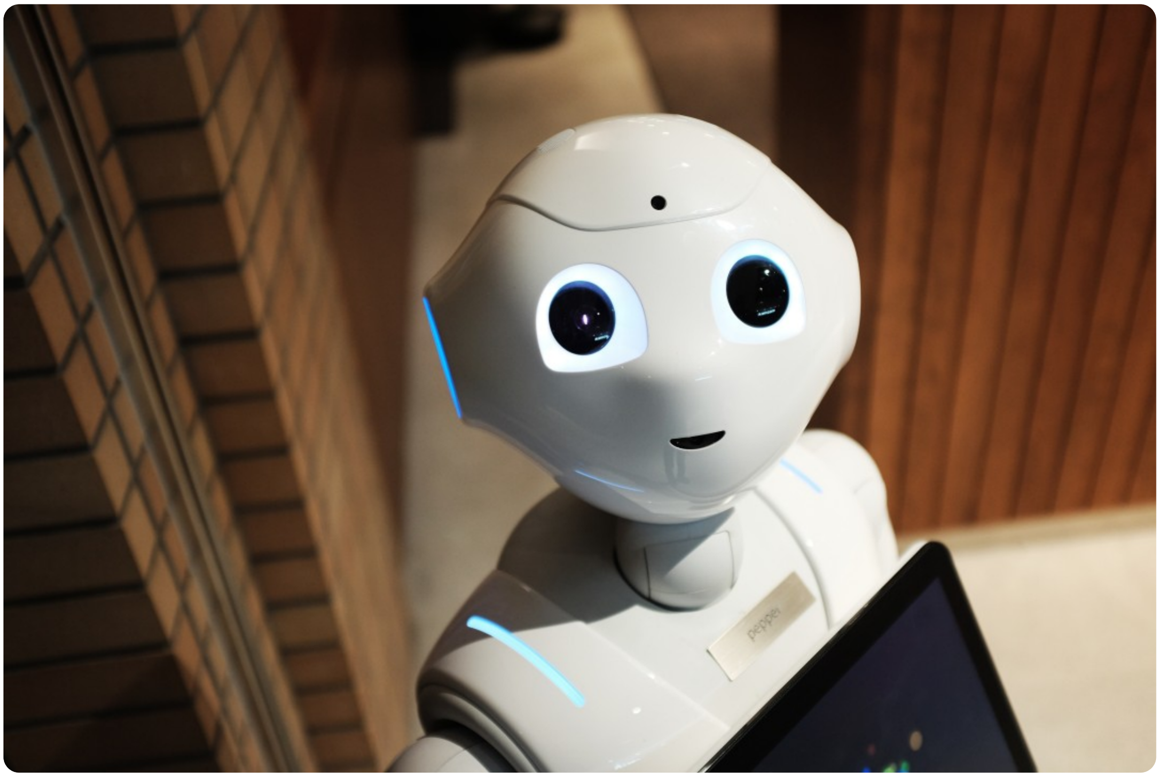

Photo by Alex Knight from Pexels

Céline Rodríguez
BI Consultant en HSO

3 articles + Follow

Microsoft has developed this set of actions to help organizations to implement a responsible AI

Fairness

system by using questions such as, how is the system intended to work? Who is the system designed to work for? Will it work for everyone equally? How can it harm others?

- **Attract a diverse pool of talent.** Ensure the design team reflects the world in which we live by including team members that have different backgrounds, experiences, education and perspectives.

- **Identify bias in datasets** by evaluating where the data came from, understanding how it was organized, and testing to ensure it is represented. Bias can be introduced at every stage in creation, from collection to modeling to operation.

- **Identify bias in machine learning algorithms** by leveraging tools and techniques that improve the transparency and intelligibility of models. Examples of these tools can be found in the next unit.

- **Leverage human review and domain expertise.** Train employees to understand the meaning and implications of AI results to ensure that they are ultimately accountable for decisions that leverage AI, especially when AI is used to inform consequential decisions about people. Finally, include relevant subject matter experts (such as those with consumer credit expertise for a credit scoring AI system) in the design process and in deployment decisions.

- **Research and employ best practices, analytical techniques, and tools** from other institutions and enterprises to help detect, prevent, and address bias in AI systems.

Reliability and Safety

Microsoft's AI Ready Assessment accessible from the link in the resources section. Use the results to determine which AI technologies will fit your organization's current maturity level and how your organization can best take advantage of AI.

- **Develop processes for auditing AI systems** in order to evaluate the quality and suitability of data and models, monitor ongoing performance, and verify that systems are behaving as intended based on established performance measures.

- **Provide detailed explanation of system operation** including design specifications, information about training data, training failures that occurred and potential inadequacies with trainings data, and the inferences and significant predictions generated.

- **Design for unintended circumstances** such as accidental system interactions, the introduction of malicious data, or cyberattacks.

- **Involve domain experts in the design and implementation processes**, especially when AI is being used to help make consequential decisions about people.

- **Conduct rigorous testing during AI system development and deployment** to ensure that systems can respond safely to unanticipated circumstances, don't have unexpected performance failures, and don't evolve in unexpected ways. AI systems involved in high-stakes scenarios that affect human safety or large populations should be tested both in lab and real-world scenarios.

- **Evaluate when and how an AI system should seek human input for impactful decisions or during critical situations**. Consider that is meaningful and intelligible. Design AI systems to ensure humans have the necessary level of input on highly impactful decisions.

- **Develop a robust feedback mechanism for users to report performance issues** so that they can be resolved quickly.

Privacy and Security

- **Comply with relevant data protection, privacy, and transparency laws** like GDPR or the California Privacy Act by investing resources in developing compliance technologies and processes or working with a technology leader during the development of AI systems. Develop processes to continually check that the AI systems are satisfying all aspects of these laws.

- **Design AI systems to maintain the integrity of personal data** so that they can only use personal data during the time it's required and for the defined purposes that have been shared with customers. Delete inadvertently collected personal data or data that is no longer relevant to the defined purpose.

- **Protect AI systems from bad actors** by designing AI systems in accordance with secure development and operations foundations, using role-based access, and protecting personal and confidential data that is transferred to third parties. Design AI systems to identify abnormal behaviors and to prevent manipulation and malicious attacks. Learn more about how to protect against new AI-specific security threats by reading our paper, Securing the Future accessible in the resources section of this module.

- **Design AI systems with appropriate controls** for customers to make choices about how and why their data is collected and used.

- **Ensure your AI system maintains anonymity** by de-identifying personal data.

- **Conduct privacy and security reviews** for all AI systems.

- **Research and implement industry best practices** for tracking relevant information about customer data, accessing and using that data, and auditing access and use.

Inclusiveness

- **Comply with laws regarding accessibility and inclusiveness** such as the Americans with Disabilities Act, the Communications and Video Accessibility Act, and the European Union laws and U.S. regulations that mandate the procurement of accessible technology.

- **Use the** Inclusive Design toolkit, available in the resources section of this module, to help system developers understand and address potential barriers in a product environment that could unintentionally exclude people.

- **Have people with disabilities test your systems** to help you determine whether the system can be used as intended by the broadest possible audience.

your system is accessible for people of all abilities.

Transparency

- **Share key characteristics of datasets** to help developers understand if a specific dataset is appropriate for their use case. For more information on tools and techniques for increasing transparency, please see the next unit, Governance and external engagements.

- **Improve model intelligibility** by leveraging simpler models and generating intelligible explanations of the model's behavior. Techniques to simplify models without sacrificing accuracy and tools to generate explanations of model's behaviors can be found in the next unit.

- **Train employees on how to interpret AI outputs** and ensure that they remain accountable for making consequential decisions based on the results.

Accountability

- **Set up internal review boards** to provide oversight and guidance on the responsible development and deployment of AI systems.

- **Ensure your employees are trained** to use and maintain the solution in a responsible and ethical manner and understand when the solution may require additional technical support.

- **Keep humans with requisite expertise in the loop** by reporting to them and involving them in decisions about model execution.

inspect, identify, and resolve challenges with model output and execution.

- **Put in place a clear system of accountability and governance** to conduct remediation or correction activities if models are seen as behaving in an unfair or potentially harmful manner.

Resources

- Understand your organization's AI Maturity by taking Microsoft's **AI Ready Assessment**.

Security and Privacy

- **Securing the Future of Artificial Intelligence and Machine Learning at Microsoft** provides guidance on how to protect algorithms, data, and services from new AI-specific security threats. While security is a constantly changing field, this paper outlines emerging engineering challenges and shares initial thoughts on potential remediation.

- Homomorphic encryption is a special type of encryption technique that allows users to compute on encrypted data without decrypting it. The results of the computations are encrypted and can be revealed only by the owner of the decryption key. To further the use of this important encryption technique, we developed the **Simple Encrypted Arithmetic Library** (SEAL) and made it open source.

encrypted data and algorithms with each other while preserving input privacy and ensuring that no party sees information about other members. For example, with MPC we can build a system that analyzes data from all three hospitals without any of them gaining access to each other's health data.

Transparency and fairness

- **Datasheets for datasets** is a paper that proposes information that dataset creators should include in a datasheet for their dataset, such as training datasets, model inputs and outputs, and model features. Like a datasheet for electronic components, a datasheet for datasets would help developers understand if a specific dataset is appropriate for their use case.

- **Local Interpretable Model-agnostics Explanations (LIME)** provides an easily understood description of a machine learning classifier by perturbing the input and seeing how the predictions change.

- **Methodology for reducing bias in word embedding** helps reduce gender biases by modifying embeddings to remove gender stereotypes, such as the association between receptionist and female, while maintaining potentially useful associations such as the association between the words queen and female.

- **A reductions approach to fair classification** provides a method for turning any common classifier into a "fair" classifier according to any of a wide range of fairness definitions. For example, consider a machine learning system tasked with choosing that predicts who should be interviewed based on previous hiring decisions into a classifier that predicts who should be interviewed while also respecting demographic parity (or another fairness definition).

- **Generalized Additive Models plus interactions (GA2M)** is a learning method based on generalized additive models that improves the transparency and intelligibility of a model without sacrificing its accuracy. By leveraging GA2M models, users can better understand what the models have learned and more easily remove bias and other errors that may have been introduced in the learning process.

Leverage principles and guides

- Use the **Inclusive Design toolkit** to help system developers understand and address potential barriers in a product environment that could unintentionally exclude people.

- **Inclusive design practices** can help system developers understand and address potential barriers in a product environment that could unintentionally exclude people.

- Guide facial recognition work with these **six principles**.

- Discuss the need for **public regulation and corporate responsibility of facial recognition technology.**

- Reference this **methodology for reducing gender bias in word embedding.**

better prepare your AI systems for these new privacy and security threats in **Securing the Future of Artificial Intelligence and Machine Learning at Microsoft.**

- Design bots that adhere to ethical principles by following these **ten guidelines.**

- Benefit from 150+ design recommendations our researchers documented into a unified set of guidelines for **human-AI interaction** to help developers design human-centered AI systems.

- **Partnership on AI** (PAI) is a group of researchers, non-profits, non-governmental organizations (NGOs), and companies dedicated to ensuring that AI is developed and utilized in a responsible manner.

Skill up

- **2018 WEF Future of Jobs Report** states many companies have been focusing their upskilling and retraining efforts on those people who already have higher skills and value to the company.

- The **Microsoft Professional Program** now has an AI track bringing together expert instructors, provide hands-on labs, offer AI-specific online courses and instructional videos.

- Developer-focused **AI School**, which provides online videos and other assets that help build professional AI skills.

- The **Skillful Initiative**, a partnership with the Markle Foundation in the US, helps match people with employers and fill high-demand

Microsoft Programs

- AI for Good includes three programs: **AI for Accessibility**, **AI for Earth**, and **AI for Humanitarian Action**, which are already supporting nearly 250 projects across the globe. Learn more about how to protect against new AI-specific security threats by reading our paper, **Securing the Future of Artificial Intelligence and Machine Learning at Microsoft** and follow along with the news. For example, last year we publicly **called for regulation** of facial recognition technology and outlined our recommendations for the public and private sector alike.

Report this

Published by

Céline Rodríguez
BI Consultant en HSO
Published • 2y

3 articles + Follow

Microsoft has developed this set of actions to help organizations to implement a responsible AI.
#ResponsibleAI #AI #Governance

Like Comment Share 7

Reactions

0 Comments

Add a comment...