Microsoft

Establishing responsible design principles in AI engineering

If you're developing, implementing, or managing AI internally, you may want to consider how to honor your organization's guiding principles at every step of the AI lifecycle. To empower your technical employees to implement your principles from design and data collection to development and deployment, we have found that it helps to translate your principles into actionable guidance and tools.

Microsoft is on this journey as well, and we would like to share our perspective and experiences. For years, we have been working with other companies and institutions to help developers everywhere build and deploy AI responsibly. We also leverage open-source tools and look to leading organizations like Partnership on AI (PAI) for best practices, industry standards, and guidelines. By leveraging practical guidance and innovative tools, hopefully you and your team don't have to develop your approach from scratch.

Principles and guidelines

Microsoft's responsible AI journey began when we established six key principles to guide our development and use of AI, which are outlined in The Future Computed: fairness, reliability and safety, privacy and security, inclusiveness, transparency, and accountability.

With these foundational principles in place, we began developing more scenario-specific guidelines. For example, in May of 2019, we published a paper called Guidelines for Human-AI Interaction, which includes 18 generally applicable design guidelines to help developers design responsible and human-centered AI systems. In addition to this key resource, we have published a number of other guidelines and principles including the ones below:

1. Design bots based on ethical principles by reviewing these ten guidelines.
2. Join Partnership on AI (PAI), 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.
3. When working with Facial Recognition, understand current and future regulation, follow a principled approach, and understand the design scenarios and limitations.

Tools

Along with our foundational principles and guidelines, we have also found a need for tools and resources that make it easier for developers to spot and design against potentially harmful issues like biases, safety and privacy gaps, and exclusionary practices.

Here are some tools that we have found helpful:

Security and privacy

Security and privacy are key pillars of trust. To win the confidence of your customers and stakeholders, use the following resources to help protect security and privacy:

1) 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.

2) Secure execution environments such as Azure confidential computing help users secure data while it's "in use" on public cloud platforms (a state required for efficient processing). The data is protected inside a Trusted Execution Environment (TEE), also known as an enclave, such that code and data are protected against viewing and modification from outside of the TEE. This has a number of benefits, including the ability to train AI models using data sources from different organizations without sacrificing data confidentiality.

    a. The Azure team has worked with Microsoft Research, Intel, Windows, and our Developer Tools group to develop our confidential computing solution, which enables developers to take advantage of different TEEs without having to change their code.

    b. The Open Enclave SDK project provides a consistent API surface for developing apps using enclave-based computing.

3) 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 Microsoft SEAL and made it open-source.

4) Multi-party computation (MPC) allows a set of parties to share 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 three different hospitals without any of them gaining access to each other's health data.

5) Differential privacy is a key technology for training machine learning models using private data. A differentially private algorithm uses random noise to ensure that the model output doesn't noticeably change when one individual in the dataset changes. This prevents attackers from inferring an individual's private information from the model's output.

    a. The PSI (Private data Sharing Interface) tool, developed by Harvard researchers, leverages differential privacy to enable researchers from many fields to explore and share datasets that contain private information.

Fairness

AI systems should treat everyone fairly and avoid affecting similarly situated groups of people in different ways. There are two key steps for achieving this—assessment and mitigation:

Assessing fairness

1) Fairness in Machine Learning (ML) Systems (FairLearn) is an approach created by Microsoft Research and co-developed with products teams. FairLearn can be used to assess the potential unfairness of ML systems that make decisions about allocating resources, opportunities, or information. Note that fairness is a fundamentally sociotechnical challenge, so "fair" classification tools are not be-all-and-end-all solutions, and they are only appropriate in particular, limited, circumstances. A Python package that implements this approach is available on GitHub.

    a. For example, consider a ML system tasked with choosing applicants to interview for a job. FairLearn can turn a classifier 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).

2) Aequitas is an open-source bias audit toolkit developed by the Center for Data Science and Public Policy at the University of Chicago for machine learning developers, analysts, and policymakers to audit machine learning models for discrimination and bias, and make informed and equitable decisions around developing and deploying predictive risk-assessment tools.

3) To understand the unique challenges regarding fairness in ML, watch our free webinar on Machine Learning and Fairness. In this webinar you'll learn how to make detecting and mitigating biases a first-order priority in your development and deployment of ML systems.

4) For more on how organizations should approach assessing the fairness of their AI models, watch this NIPS keynote address from Kate Crawford, Principle Researcher at Microsoft and Co-founder of the AI Now Institute at NYU.

Mitigating bias

1) A methodology for reducing bias in word embedding created by Microsoft Research helps reduce gender biases by modifying embeddings to remove gender stereotypes, such as the association between receptionist and female.

2) Read this paper from the ACM Conference on Fairness, Accountability, and Transparency: Fairness and Abstraction in Sociotechnical Systems which explains five key "traps" of fair-ML work and how to avoid them.

3) Read this paper from Cornell University: Counterfactual Fairness for an example of a framework for modeling fairness using tools from causal interference, and how it applies to the fair prediction of student success in law school.

Inclusiveness

Inclusive design practices help ensure that AI models perform well for all users and no one is excluded from the opportunities provided by intelligent solutions. To help address potential barriers in your product environment that could unintentionally exclude people, use the following resources:

1) The Microsoft Research paper Algorithmic Greenlining proposes an approach that helps decision-makers develop selection criteria yielding high-quality and diverse results in contexts such as college admissions, hiring, and image search.

    a. Take, for example, choosing job candidate search criteria. There's typically limited information about any candidate's "true quality." An employer's intuition might suggest searching for "computer programmer," which yields high-quality candidates but might return few female candidates. The algorithmic framework suggests alternative queries which are similar but more gender-diverse, such as "software developer" or "web developer."

2) Reference Microsoft's inclusive design toolkit and inclusive design practices to learn how to understand and address potential barriers in a product environment that could unintentionally exclude people.

Reliability and safety

AI systems can become unreliable or inaccurate if their development and testing environment is not the same as the real world or if the system is not maintained properly. This can be especially dangerous in industries where safety may be at risk, like manufacturing or healthcare. To prevent reliability and safety issues, there are a number of technologies and tools that strengthen model performance through long-term monitoring and management:

1) The Data Drift Monitoring feature in Azure Machine Learning detects changes in the distribution of data that may cause degraded prediction performance, enabling developers to maintain accuracy by adapting the model to reflect changing data.

2) Pandora is a debugging framework designed by Microsoft Research to identify reliability and bias problems within machine learning models. It uses interpretable machine learning techniques, such as decision trees, to discover patterns and identify potential issues.

3) Microsoft AirSim is a valuable open-source tool for improving simulated training environments.

Transparency

The black-box nature of AI can be problematic and potentially harmful. To help your organization articulate how your AI models reach conclusions and build trust with your users, use the following resources:

1) InterpretML is an open-source package created by Microsoft Research for training interpretable models and explaining black box systems. It implements a number of intelligible models including Explainable Boosting Machine (EBM), an improvement over generalized additive models that has both high accuracy and intelligibility. It also supports several methods for generating explanations of black box model behavior or predictions including 'SHapley Additive exPlanations' (SHAP) and 'Local Interpretable Model-agnostics Explanations' (LIME).

2) Azure Machine Learning has a variety of tools that support model transparency. The Model Interpretability feature enables model designers and evaluators to explain why a model makes the predictions it does, which can be used to debug the model, validate that its behavior matches objectives, and check for bias.

Accountability

The people who design and deploy AI systems must be accountable for how their systems operate. A useful first step for developing accountability and transparency in your organization is to create thorough documentation processes for AI systems. To develop accountability practices for your own organization, leverage the following resources:

1) Datasheets for datasets is a paper that encourages people assembling training datasets to generate a datasheet with key information such as the motivation, composition, collection process, and recommended uses. Datasheets for datasets have the potential to increase transparency and accountability within the machine learning community, mitigate unwanted biases in machine learning systems, facilitate greater reproducibility of machine learning results, and help researchers and practitioners select more appropriate datasets for their chosen tasks.

2) The DevOps feature in Azure Machine Learning (called MLOps) makes it easier to track, reproduce, and share models and their version histories. It offers centralized management throughout the entire model development process, and helps teams monitor model performance by collecting application and model telemetry.

3) The Partnership on AI (PAI) is leading a multi-stakeholder initiative called ABOUT ML to develop, test, and promulgate best practices for machine learning documentation. These best practices may include documenting how AI systems were designed and for what purposes, where their data came from and why that data was chosen, how they were trained, tested, and corrected, and what purposes they're not suitable for.