PERSPECTIVES FROM THE FRONT

Protecting individuals and enterprises through algorithmic safety

Data & Trust Alliance

IBM

Algorithmic safety refers to processes that ensure algorithms don't produce biased or harmful results. Algorithmic safety practices include evaluating the quality of training data, ensuring that the algorithms are appropriate for the context and purpose for which they're used, and providing education and training for the people who build and use AI tools.

THE SECURITY AND ETHICS of AI have become urgent concerns as the technology becomes embedded in the operations of many, if not most, organizations — from business to government to education. Given the value that AI provides, and the disruptive opportunities it presents, its use will only become more pervasive.

A major barrier to AI-driven business progress, however, is the real harm that these tools can cause. Take human resources — on one hand, AI-based tools can help enterprises diversify talent pools and effectively match candidates' skills to workplace opportunities. Since the 1990s, the recruitment industry has used increasing amounts of automation, and the technology is pervasive today among employers and third-party service providers such as LinkedIn, Indeed and ZipRecruiter. In Eightfold AI's 2022 report The Future of Work: Intelligent by Design, 78% of the 250 HR leaders surveyed said they are already using AI across HR functions like employee records management. Seventy-seven percent are using AI for payroll processing and benefits administration, 73% for recruitment and hiring, 72% for performance management and 69% for onboarding new employees.

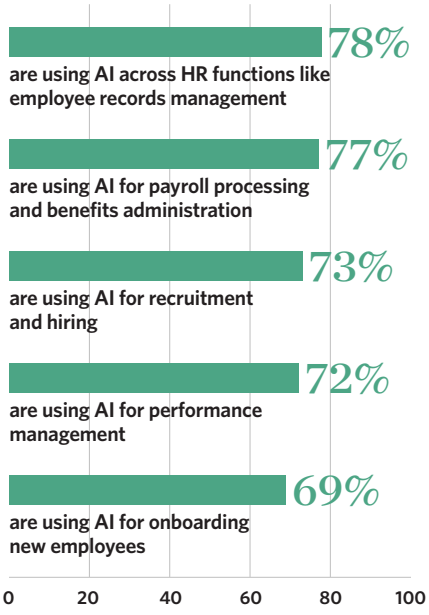
However, if those tools are trained on data sets that reflect historical biases in companies' hiring and promotion practices, the tools will perpetuate the same bias. The same applies to AI tools for financial services, which could propagate historical biases that impact lending.

Walmart uses AI not only for HR processes but also to ensure the freshness of packaged food products, and to help delivery drivers plot routes that are faster, safer and more fuel efficient. As America's second-largest employer (after the federal government), Walmart's priorities for algorithmic safety are focused on the outputs of AI models that directly affect people. "In particular, we're focused on use cases where AI tools and technologies impact someone's rights or status in the world," says Nuala O'Connor, Walmart's senior vice president and chief counsel for digital citizenship. "Most significant for us are HR and hiring, advancement, promotion, education and training opportunities."

Fairness is inarguably a proper goal for all organizations, and when it comes to AI decisioning, it's increasingly becoming the law as governments worldwide move toward regulating AI. In July 2023, New York City began enforcing a landmark law that requires employers using AI and other algorithms for recruiting, hiring or promotion to have their tools audited for bias by a third party.

Some 78% of business and IT decision makers surveyed by Progress, a provider of application development and infrastructure

How 250 HR leaders surveyed are using AI



Source: The Future of Work: Intelligent by Design

software, say they believe bias will become a bigger concern as AI use increases. However, only 13% say they are currently addressing it and have an evaluation process in place.

The following five steps to implementing algorithmic safety frameworks will enable business leaders to take a pragmatic, human-centered approach to their use of AI tools and technologies.

STEP ONE

Define algorithmic safety for your organization

There is no one-size-fits-all approach to algorithmic safety. Every organization should start by creating a definition of algorithmic safety that reflects its strategy and principles, as well as the specific use cases for its AI tools.

Walmart built its algorithmic safety practices around the company's four core values: serving the customer, respecting the individual, striving for excellence and acting with integrity. "One of the most important things we did was to take our four core values and reimagine them in ways that resonate with our uses of technology and data," O'Connor says. Across all use cases, O'Connor says, starting with a values-based foundation for AI means "creating tools that are going to elevate the dignity of the individual by treating them fairly; treating them with respect."

A strong algorithmic safety practice should start with the value an organization aims to provide to its stakeholders, from employees to partners to consumers. Algorithmic safety then becomes a matter of ensuring that AI tools treat people the same way the enterprise aspires to treat people.

STEP TWO

Understand the dimensions of algorithmic safety

Algorithmic safety encompasses more than bias. It is also about ensuring that AI models are trustworthy, which can be compromised by three problems:

- Bad quality in the data used to train the algorithm — for example, if the data is misaligned with the people or situation for which the model is being used
- Technical flaws in the algorithms themselves
- Human error — both in the training and deployment of the algorithms, and in how humans eventually use AI tools

Any one of these problems can result in AI tools that generate untrustworthy outcomes — and that, in turn, can lead to bad business practices, as well as legal and compliance risks. For that reason, algorithmic safety efforts need to target both the creation and use of AI models.

> "One of the most important things we did was to take our four core values and reimagine them in ways that resonate with our uses of technology and data."
>
> —**NUALA O'CONNOR,** senior vice president and chief counsel for digital citizenship at Walmart

Many providers of AI technologies have frameworks for mitigating biases while training AI models. These include IBM's AI Fairness 360 Toolkit, Microsoft's Fairlearn, Meta's Casual Conversations v2 and Google's TensorFlow Model Remediation library.

STEP THREE

Implement a screening process

A core component of an algorithmic safety framework is comprehensive, standardized screening processes for all AI technologies — those provided by vendors as well as those built in-house. Screening must evaluate the quality of the data used to train AI

models, as well as the implicit biases of the people and organizations who created the algorithms and tools. According to business and IT leaders surveyed by Progress, one of the biggest barriers to mitigating risk in AI tools is the difficulty of identifying those biases.

Screening should also assess whether the tool is fit for the purpose for which it will be used. Even a tool that is responsibly built can produce untrustworthy outcomes if it's deployed for purposes or in contexts that are not aligned with the function the tool was built to perform. It's important to have a mechanism in place for vetting use cases, as well as for evaluating the way a model is built and developed.

For most companies, a relatively simple questionnaire may be sufficient for the screening process. O'Connor worked with the Data & Trust Alliance to create safeguards specifically for HR tools. They include a 55-question survey for third-party vendors and a scorecard to help compare vendors. Her own team uses a questionnaire that assesses key issues:

- What does the tool do?
- What is it for, and who is going to use it?
- What data has been used to build the tool?
- What outcomes are being generated?
- How is the tool going to be implemented in the real world?
- Who will have oversight, and how will they be trained?

In most cases, a red flag in this evaluation simply offers an opportunity for an engineer or a product lead to meet with the compliance team and discuss the tool in detail.

Evaluation doesn't end when the product launches — AI model drift is among the biggest risks for AI tools. Organizations need to be able to document the origin model, explain how and why it generates the outcomes it does, and track its changes over time. For this reason, algorithmic safety should ultimately be part of a larger AI governance program.

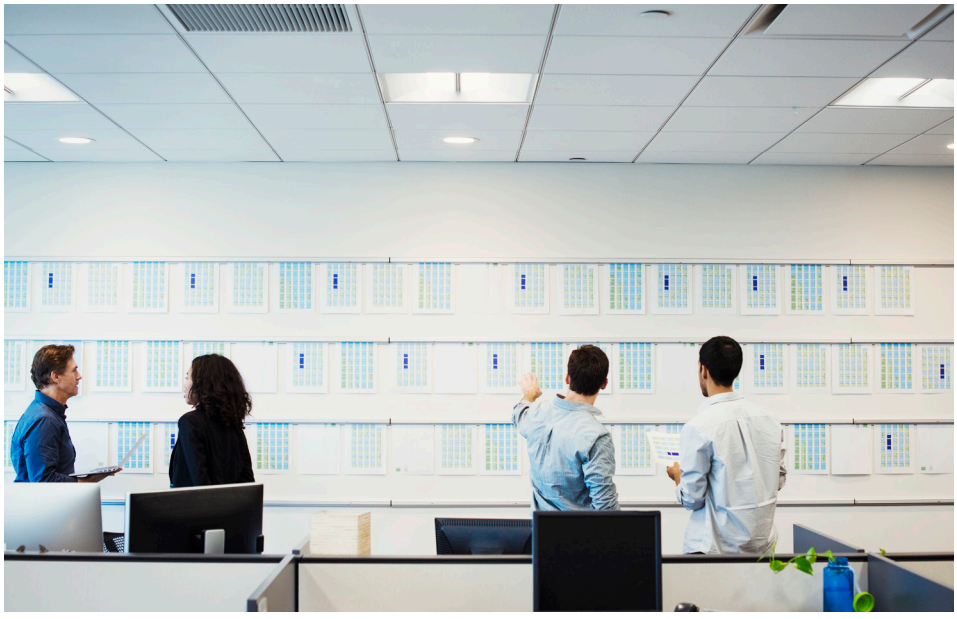
STEP FOUR

Invest in communication, education and training

Most fundamentally, algorithmic safety is a question of how humans interact with AI. Therefore, training and education are critical.

"There are two different types of training required," says Dena Mendelsohn, privacy officer and compliance manager at health-tech startup Transcarent. "You need to train the designers and builders of the AI models to be aware of the different risks, controls and quality assurance requirements both pre- and post-launch. And if your workforce uses the output of the models, they need to be trained on proper use of the model and how to avoid overreliance or use that is out of scope of the original design."

For end users, understanding how to think critically about the outcomes of an AI model —



"You really need to have training to be able to interpret what you're getting from a model."

—**CHRISTINE PIERCE,** chief data officer at Nielsen

rather than blindly trusting the results — is an often overlooked yet crucial aspect of algorithmic safety. "You really need to have training to be able to interpret what you're getting from a model," says Christine Pierce, chief data officer at Nielsen. For instance, Pierce says, "If a medical professional is used to giving a certain type of diagnosis, and now they're getting information from an AI tool that gives them a range of probability, they need training to know what to do with that."

Organizations should communicate their algorithmic safety values throughout the workforce and provide ongoing, targeted training for both the teams that build and deploy AI tools and the people who use them. O'Connor says that Walmart offers online training modules as well as in-person events, and regularly discusses algorithmic safety on company-wide communications channels like Walmart Radio.

STEP FIVE

Operationalize algorithmic safety as a compliance program

While algorithmic safety is a new concern, putting it into practice is similar to implementing many existing risk-related compliance programs, such as workplace safety. Organizations can adapt core elements of these compliance programs, including continuous performance tracking, logging incidents and recording both failures and near misses as well as remediation strategies. A classic compliance approach to AI governance might include registering all AI models, regularly testing for bias using a standardized screening process, documenting all results and putting controls in place to remediate any bias that is detected.

Rather than being a hindrance to technological evolution, algorithmic safety can help drive technological innovation at speed and at scale. "The challenge, for those of us who work in what I would call tech-adjacent roles, is to move fast alongside our tech partners but still make sure we are asking the right questions to keep the boundaries really clear about what's okay and what's not," O'Connor says. "You can appreciate and get excited about technology advancements without losing sight of necessary boundaries. They are making our associates' and our customers' lives better." ●

Key questions for the C-suite

- Do we have a **clear definition of algorithmic safety** for our critical stakeholders, grounded in our company's mission and values?
- Do we have a structured, cross-enterprise **screening process** for algorithmic safety?
- Do we provide **algorithmic safety training** for both builders of AI models and their users?
- How do we communicate and operationalize a **culture of algorithmic safety** company-wide?
- Is algorithmic safety **integrated into our enterprise compliance processes?**

→ Learn more about the Data & Trust Alliance

Get expert insights on AI for business