HARVARD
BUSINESS
REVIEW

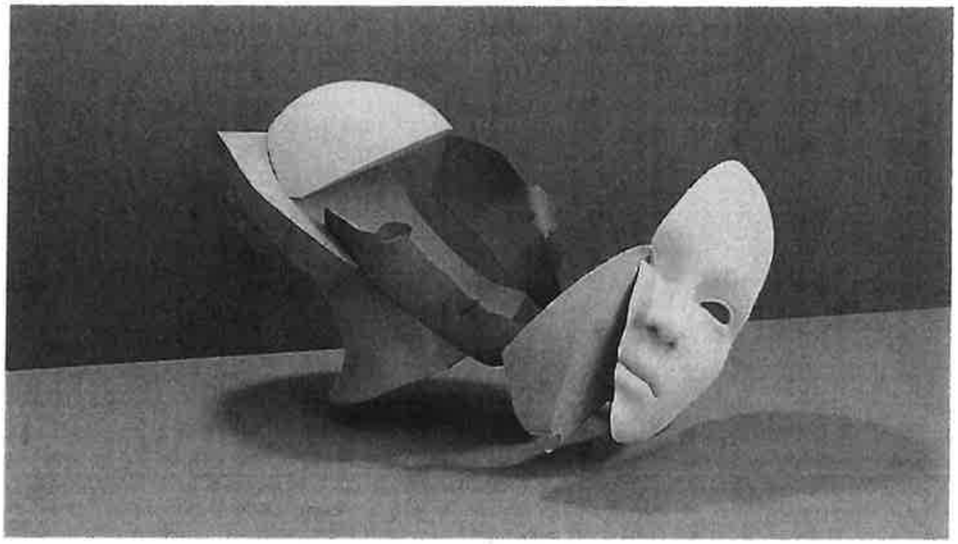
Digital Article / AI and Machine Learning

Inside Adobe's Approach to Assessing AI Risk

Four principles the company uses to preempt legal and ethical trouble.

by Grace Yee

Published on HBR.org / May 9, 2025 / Reprint H08Q76



Hector Roqueta Rivero/Getty Images

While concerns about the risks of generative AI persist, people are increasingly embracing AI's power and practical applications in everyday life. But despite sharing common underlying technology, AI applications range from the mundane to the complex. This broad spectrum of use cases makes it challenging to determine which guardrails are appropriate to support responsible AI innovation.

To do this effectively, it's critical that leaders carefully consider the context in which AI operates—understanding the specific circumstances, environments, and applications of AI technologies in real-world use cases. AI is often assessed through its inputs and outputs, but this paints only a partial picture. By prioritizing contextual analysis, companies can design safeguards that help anticipate and mitigate risks across a wide range of scenarios and use cases.

Companies need to shift from viewing AI as a singular entity to evaluating its impact in specific contexts. Here are four principles to consider while codifying a context-driven ethical review process that will allow for more precise risk assessment and ensure safeguards align with real-world use cases.

Human involvement is critical.

Different AI applications vary widely in their level of impact. The same large language model (LLM) that generates an inspirational travel recommendation might also fabricate critical medical information—yet only one has life-altering consequences. While these LLMs may share technical similarities, their risks differ drastically to the person using that technology.

And with the rise of agentic AI, leaders need to stay agile with how they assess use cases. According to Deloitte, by 2027, half of companies that use generative AI will have launched agentic AI pilots that can act as smart assistants, performing complex tasks with minimal human supervision. While greater autonomy might enable AI agents to streamline operations or enhance customer interactions, without appropriate human oversight it can also lead to serious errors. When deploying agentic AI, maintaining human oversight will be essential to ensuring that AI agents do not take actions on users' behalf without their awareness.

At Adobe, we assess AI features by categorizing them as low or high impact, prioritizing thoughtful review of the highest-risk areas. While AI risks exist on a spectrum, this distinction helps us focus on where harm is most significant. For instance, an AI model that recommends font style carries little downside—at worst, it provides poor suggestions. But a text-to-image generation model has much broader, unpredictable implications.

By applying high-risk/low-risk analysis to AI governance, leaders can ensure that human oversight remains a central safeguard as AI evolves. This focused approach allows organizations to mitigate critical risks while continuing to foster innovation.

Hype doesn't equal quality.

In the AI race, pushing out products quickly has often taken priority over ensuring reliability and accuracy and mitigating harms. A flashy AI demo may seem impressive, but without rigorous testing, it can lack real-world effectiveness and even cause real-world harm. Innovation should never compromise on quality or perspective—bringing a range of perspectives into early AI development and continuously testing the technology can help ensure it's used as intended in real-world contexts.

There's a reason companies invest in beta testing: Internal testing alone can't predict every way users will engage with AI. Real-world feedback is essential for identifying unintended behaviors and improving reliability. If you're a global company or aspire for your technology to become global, you must expand the pool of people testing and using your technology to account for broad scenarios and cultural use cases.

For example, when developing Adobe Firefly, our family of creative generative AI models, we initially restricted the text-to-image model from generating content based on the word "skeleton" as a precaution against potentially harmful or disturbing outputs. However, through user feedback, we realized this restriction also blocked harmless, everyday use cases, like generating skeleton images for Halloween. Without real-world input, we could have unintentionally limited creative expression and valuable applications of the technology.

AI guardrails aren't one-size-fits-all—they must be adaptable. What's necessary in one context may be overly restrictive in another. Ongoing feedback from a diverse range of users helps ensure safeguards stay effective and relevant.

AI will always evolve, but you can anticipate concerns.

A common misconception is that safer AI models inherently lead to better outcomes. In reality, there's no such thing as completely safe AI. You can design a model to be "safe," yet it can still cause unintended harm.

Future-proofing AI isn't just about avoiding harmful outputs—it's about building AI that remains viable, ethical, and legally sound over time. AI companies are now facing copyright lawsuits, highlighting the risks of training models on unlicensed data. If businesses embed AI into their products without a stable foundation, they risk costly legal battles, retraining efforts, or even having to strip the technology altogether from their services.

Understanding a model's training data is vital for assessing the limitations and potential risks. Our Firefly generative AI models are commercially safe and IP-friendly because they're exclusively trained on content Adobe has permission to use, including licensed Adobe Stock content and public domain content. But that doesn't mean we don't rigorously test for risks—we still do. Guardrails and safety mechanisms should be context-aware and tailored to the model's training data.

Leaders should take a proactive approach by embedding legal and ethical risk assessment early in AI development, rather than addressing issues retroactively. By rigorously testing models in real-world scenarios and adapting safeguards to evolving risks, businesses can create AI that not only meets today's standards but also remains resilient against future challenges.

AI requires a robust ethics framework.

Responsible innovation begins with a strong ethical foundation. While AI is complex, your review process for the technology must be clear and actionable. To build an effective AI review and governance framework, leaders must set the vision and ensure AI strategy aligns with company values.

At Adobe, our leadership team convened a cross-functional group six years ago to help establish our AI ethics principles of Accountability, Responsibility, and Transparency as the foundation for our approach to developing AI responsibly. From this cross-functional group emerged the AI Ethics Impact Assessment, designed to help apply these principles consistently across our AI innovations. This assessment examines key considerations for new AI features we introduce to our customers, including who will use them and how; what safeguards ensure the AI model is safe, reliable, and aligned with societal needs; and how intended and unintended harms will be assessed and mitigated. To ensure its effectiveness, we piloted it with product teams first, refining the questions based on their feedback to ensure they were practical, relevant, and actionable before scaling it company-wide. By embedding the AI Ethics Impact Assessment into existing workflows, we seamlessly integrate it into the product development process.

An AI ethics team is also essential for ensuring that innovation is paired with responsible innovation. Leaders should empower this team to engage deeply with product teams, conducting proactive risk discovery exercises to identify and mitigate against potential harms. Through this approach, our AI ethics team is able to gain a foundational understanding of how new AI features work as well as their intended use case and target audience, allowing us to identify real-world risks through rigorous hands-on testing.

Beyond launch, leadership support allows the AI ethics team to play a critical role in ongoing testing, mitigating bias and improving usability. However, ethical AI development cannot rest on this one team alone—it must be a shared responsibility across the company. Leaders must champion this approach to ensure AI remains both cutting-edge and responsible.

■ ■ ■

AI ethics cannot rely on a rigid, one-size-fits-all approach. Responsible AI development and deployment requires understanding context, assessing impact, and continuously refining approaches. A nuanced, context-driven ethics framework helps ensure AI is deployed responsibly while fostering innovation.

Furthermore, incorporating a broad range of perspectives helps keep AI aligned with societal needs, minimizing potential harm across industries and applications. Industry, governments, and users benefit most from AI when all parties recognize that assessing AI systems in context is the most effective way to address challenges, while driving progress and efficiency gains.

Without this focus, AI risks becoming detached from real-world needs, amplifying harm instead of reducing it. Prioritizing context-driven ethics ensures AI remains not just innovative but also reliable, responsible, and ultimately trustworthy.

This article was originally published online on May 9, 2025.

Grace Yee is the Senior Director of Ethical Innovation at Adobe.