AI And Machine Learning

Managing the Risks of Generative AI

by Kathy Baxter and Yoav Schlesinger

June 6, 2023

serggn/Getty Images

Summary. Generative artificial intelligence (AI) has become widely popular, but its adoption by businesses comes with a degree of ethical risk. Organizations must prioritize the responsible use of generative AI by ensuring it is accurate, safe, honest, empowering,... **more**

Corporate leaders, academics, policymakers, and countless others are looking for ways to harness generative AI technology, which has the potential to transform the way we learn, work, and more. In business, generative AI has the potential to transform the way companies interact with customers and drive business growth. New research shows 67% of senior IT leaders are prioritizing generative AI for their business within the next 18 months, with one-third (33%) naming it as a top priority. Companies are exploring how it could impact every part of the business, including sales, customer service, marketing, commerce, IT, legal, HR, and others.

However, senior IT leaders need a trusted, data-secure way for their employees to use these technologies. Seventy-nine-percent of senior IT leaders reported concerns that these technologies bring the potential for security risks, and another 73% are concerned about biased outcomes. More broadly, organizations must recognize the need to ensure the ethical, transparent, and responsible use of these technologies.

A business using generative AI technology in an enterprise setting is different from consumers using it for private, individual use. Businesses need to adhere to regulations relevant to their respective industries (think: healthcare), and there's a minefield of legal, financial, and ethical implications if the content generated is inaccurate, inaccessible, or offensive. For example, the risk of harm when an generative AI chatbot gives incorrect steps for cooking a recipe is much lower than when giving a field service worker instructions for repairing a piece of heavy machinery. If not designed and deployed with clear ethical guidelines, generative AI can have unintended consequences and potentially cause real harm.

Organizations need a clear and actionable framework for how to use generative AI and to align their generative AI goals with their businesses' "jobs to be done," including how generative AI will impact sales, marketing, commerce, service, and IT jobs.

In 2019, we published our trusted AI principles (transparency, fairness, responsibility, accountability, and reliability), meant to guide the development of ethical AI tools. These can apply to any organization investing in AI. But these principles only go so far if organizations lack an ethical AI practice to operationalize them

organizations lack an ethical AI practice to operationalize them into the development and adoption of AI technology. A mature ethical AI practice operationalizes its principles or values through responsible product development and deployment — uniting disciplines such as product management, data science, engineering, privacy, legal, user research, design, and accessibility — to mitigate the potential harms and maximize the social benefits of AI. There are models for how organizations can start, mature, and expand these practices, which provide clear roadmaps for how to build the infrastructure for ethical AI development.

But with the mainstream emergence — and accessibility — of generative AI, we recognized that organizations needed guidelines specific to the risks this specific technology presents. These guidelines don't replace our principles, but instead act as a North Star for how they can be operationalized and put into practice as businesses develop products and services that use this new technology.

Guidelines for the ethical development of generative AI

Our new set of guidelines can help organizations evaluate generative AI's risks and considerations as these tools gain mainstream adoption. They cover five focus areas.

Accuracy

Organizations need to be able to train AI models on their own data to deliver verifiable results that balance accuracy, precision, and recall (the model's ability to correctly identify positive cases within a given dataset). It's important to communicate when there is uncertainty regarding generative AI responses and enable people to validate them. This can be done by citing the sources where the model is pulling information from in order to create content, explaining why the AI gave the response it did,

highlighting uncertainty, and creating guardrails preventing some tasks from being fully automated.

Safety

Making every effort to mitigate bias, toxicity, and harmful outputs by conducting bias, explainability, and robustness assessments is always a priority in AI. Organizations must protect the privacy of any personally identifying information present in the data used for training to prevent potential harm. Further, security assessments can help organizations identify vulnerabilities that may be exploited by bad actors (e.g., "do anything now" prompt injection attacks that have been used to override ChatGPT's guardrails).

Honesty

When collecting data to train and evaluate our models, respect data provenance and ensure there is consent to use that data. This can be done by leveraging open-source and user-provided data. And, when autonomously delivering outputs, it's a necessity to be transparent that an AI has created the content. This can be done through watermarks on the content or through in-app messaging.

Empowerment

While there are some cases where it is best to fully automate processes, AI should more often play a supporting role. Today, generative AI is a great *assistant*. In industries where building trust is a top priority, such as in finance or healthcare, it's important that humans be involved in decision-making — with the help of data-driven insights that an AI model may provide — to build trust and maintain transparency. Additionally, ensure the model's outputs are accessible to all (e.g., generate ALT text to accompany images, text output is accessible to a screen reader). And of course, one must treat content contributors, creators, and

And of course, one must treat content contributors, creators, and data labelers with respect (e.g., fair wages, consent to use their work).

Sustainability

Language models are described as "large" based on the number of values or parameters it uses. Some of these large language models (LLMs) have hundreds of billions of parameters and use a lot of energy and water to train them. For example, GPT3 took 1.287 gigawatt hours or about as much electricity to power 120 U.S. homes for a year, and 700,000 liters of clean freshwater.

When considering AI models, larger doesn't always mean better. As we develop our own models, we will strive to minimize the size of our models while maximizing accuracy by training on models on large amounts of high-quality CRM data. This will help reduce the carbon footprint because less computation is required, which means less energy consumption from data centers and carbon emission.

Integrating generative AI

Most organizations will integrate generative AI tools rather than build their own. Here are some tactical tips for safely integrating generative AI in business applications to drive business results:

Use zero-party or first-party data

Companies should train generative AI tools using zero-party data — data that customers share proactively — and first-party data, which they collect directly. Strong data provenance is key to ensuring models are accurate, original, and trusted. Relying on third-party data, or information obtained from external sources, to train AI tools makes it difficult to ensure that output is accurate.

For example, data brokers may have old data, incorrectly combine data from devices or accounts that don't belong to the same person, and/or make inaccurate inferences based on the data. This applies for our customers when we are grounding the models in their data. So in Marketing Cloud, if the data in a customer's CRM all came from data brokers, the personalization may be wrong.

Keep data fresh and well-labeled

AI is only as good as the data it's trained on. Models that generate responses to customer support queries will produce inaccurate or out-of-date results if the content it is grounded in is old, incomplete, and inaccurate. This can lead to hallucinations, in which a tool confidently asserts that a falsehood is real. Training data that contains bias will result in tools that propagate bias.

Companies must review all datasets and documents that will be used to train models, and remove biased, toxic, and false elements. This process of curation is key to principles of safety and accuracy.

Ensure there's a human in the loop

Just because something can be automated doesn't mean it should be. Generative AI tools aren't always capable of understanding emotional or business context, or knowing when they're wrong or damaging.

Humans need to be involved to review outputs for accuracy, suss out bias, and ensure models are operating as intended. More broadly, generative AI should be seen as a way to augment human capabilities and empower communities, not replace or displace them.

Companies play a critical role in responsibly adopting generative AI, and integrating these tools in ways that enhance, not diminish, the working experience of their employees, and their customers. This comes back to ensuring the responsible use of AI in maintaining accuracy, safety, honesty, empowerment, and sustainability, mitigating risks, and eliminating biased outcomes. And, the commitment should extend beyond immediate corporate interests, encompassing broader societal responsibilities and ethical AI practices.

Test, test, test

Generative AI cannot operate on a set-it-and-forget-it basis — the tools need constant oversight. Companies can start by looking for ways to automate the review process by collecting metadata on AI systems and developing standard mitigations for specific risks.

Ultimately, humans also need to be involved in checking output for accuracy, bias and hallucinations. Companies can consider investing in ethical AI training for front-line engineers and managers so they're prepared to assess AI tools. If resources are constrained, they can prioritize testing models that have the most potential to cause harm.

Get feedback

Listening to employees, trusted advisors, and impacted communities is key to identifying risks and course-correcting. Companies can create a variety of pathways for employees to report concerns, such as an anonymous hotline, a mailing list, a dedicated Slack or social media channel or focus groups. Creating incentives for employees to report issues can also be effective.

Some organizations have formed ethics advisory councils

Some organizations have formed ethics advisory councils — composed of employees from across the company, external experts, or a mix of both — to weigh in on AI development. Finally, having open lines of communication with community stakeholders is key to avoiding unintended consequences.

• • •

With generative AI going mainstream, enterprises have the responsibility to ensure that they're using this technology ethically and mitigating potential harm. By committing to guidelines and having guardrails in advance, companies can ensure that the tools they deploy are accurate, safe and trusted, and that they help humans flourish.

Generative AI is evolving quickly, so the concrete steps businesses need to take will evolve over time. But sticking to a firm ethical framework can help organizations navigate this period of rapid transformation.

Kathy Baxter is Principal Architect of Ethical AI Practice at Salesforce, developing research-informed best practices to educate Salesforce employees, customers, and the industry on the development of responsible AI. She collaborates and partners with external AI and ethics experts to continuously evolve Salesforce policies, practices, and products. She is a member of Singapore's Advisory Council on the Ethical Use of AI and Data, Visiting AI Fellow at NIST, and on the Board of EqualAI. Prior to Salesforce, she worked at Google, eBay, and Oracle in User Experience Research. She is the co-author of "Understanding Your Users: A Practical Guide to User Research

Methodologies.”




Read more on **AI and machine learning** or related topics **Technology and analytics**, **Business ethics**, **Risk management**, **IT management** and **Digital transformation**

Yoav Schlesinger is an Architect of Ethical AI Practice at Salesforce, helping the company embed and instantiate ethical product practices to maximize the societal benefits of AI. Prior to coming to Salesforce, Yoav was a founding member of the Tech and Society Solutions Lab at Omidyar Network, where he launched the Responsible Computer Science Challenge and helped develop EthicalOS, a risk mitigation toolkit for product managers.