Deloitte.

Proactive risk management in Generative AI

Deloitte AI Institute

About the AI Institute

The Deloitte AI Institute helps organizations connect the different dimensions of a robust, highly dynamic and rapidly evolving AI ecosystem.

The AI Institute leads conversations on applied AI innovation across industries, with cutting-edge insights, to promote human-machine collaboration in the “Age of With”.

The Deloitte AI Institute aims to promote a dialogue and development of artificial intelligence, stimulate innovation, and examine challenges to AI implementation and ways to address them. The AI Institute collaborates with an ecosystem composed of academic research groups, start-ups, entrepreneurs, innovators, mature AI product leaders, and AI visionaries, to explore key areas of artificial intelligence including risks, policies, ethics, future of work and talent, and applied AI use cases. Combined with Deloitte’s deep knowledge and experience in artificial intelligence applications, the Institute helps make sense of this complex ecosystem, and as a result, deliver impactful perspectives to help organizations succeed by making informed AI decisions.

No matter what stage of the AI journey you’re in; whether you’re a board member or a C-Suite leader driving strategy for your organization, or a hands on data scientist, bringing an AI strategy to life, the Deloitte AI institute can help you learn more about how enterprises across the world are leveraging AI for a competitive advantage. Visit us at the Deloitte AI Institute for a full body of our work, subscribe to our podcasts and newsletter, and join us at our meet ups and live events. Let’s explore the future of AI together.

www.deloitte.com/us/AIInstitute

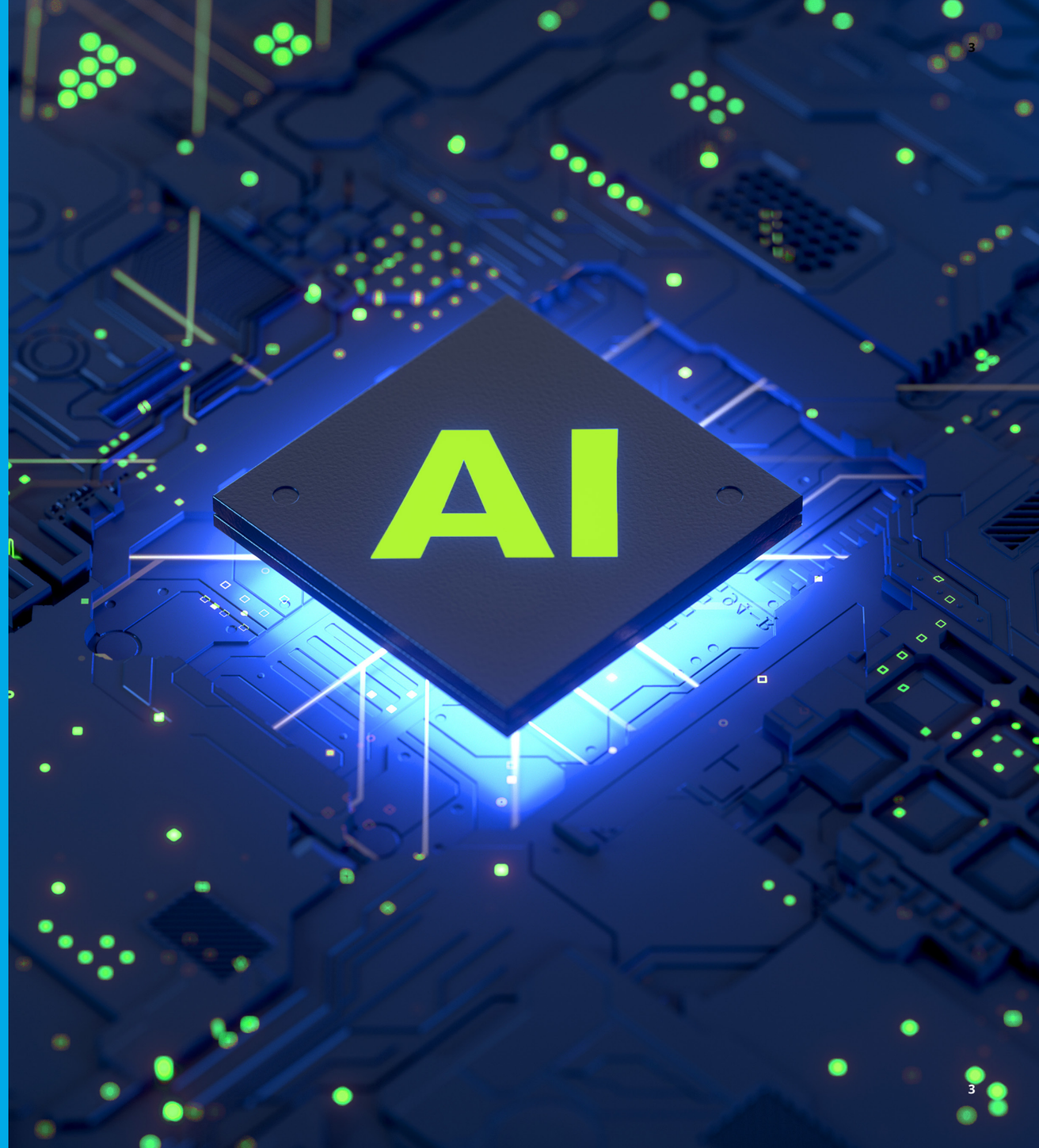
Generative AI is dominating public interest in artificial intelligence

By some estimations, Generative AI is the end of the Internet search and the tool that will revolutionize many aspects of how we work and live. We've heard that before in AI. The newest applications often conjure public excitement.

Yet, Generative AI is different than most other kinds of AI in use today. Large language models, for example, can respond to user prompts with natural language outputs that convincingly mimic coherent human language. What is more, there is effectively no barrier to using some of these models because they do not require any knowledge of AI, much less an understanding of the underlying math and technologies.

In the business realm, there is growing intrigue around how Generative AI can be used in the enterprise. As with all cognitive tools, the outcomes depend on how they are used, and that includes managing the risks, which for Generative AI have not been as deeply explored as the capabilities.

Some primary questions are, can business users trust the outputs of this kind of AI application, and if not, how can that be achieved?



New bots on the block

To this point, **AI has broadly been used to automate tasks, uncover patterns and correlations, and make accurate predictions about the future** based on current and historical data. Generative AI is designed to create data that looks like real data. Put another way, Generative AI produces digital artifacts that appear to have the same fidelity as human-created artifacts. Natural language prompts, for example, can lead the neural network to generate images that are in some cases indistinguishable from authentic images. For large language models that create text, the AI sometimes supplies source information, underscoring to the user that its outputs are factually true, as well as persuasively phrased. **"Trust me," it seems to say.**

CIOs and technologists may already know that Generative AI is not "thinking" or being creative in a human way, and they also likely know that the **outputs are not necessarily as accurate as they might appear.** Non-technical business users, however, may not know how Generative AI functions or how much confidence to place in its outputs. The business challenge is magnified by the fact that this area of AI is evolving at a rapid pace. **If organizations and end users are challenged just to keep up with Generative AI's evolving capabilities, how much more difficult might it be to anticipate the risks and enjoy real trust in these tools?**

Trust is not an inherent quality of AI but instead the product of AI governance, risk mitigation, and the intentional alignment of people, processes, and technologies across the enterprise.

The trustworthiness of Generative AI depends on how an organization uses it, and as enterprises wade into this fast-moving field of AI, there are factors of trust and ethics that should be addressed.

1 | Managing hallucinations and misinformation

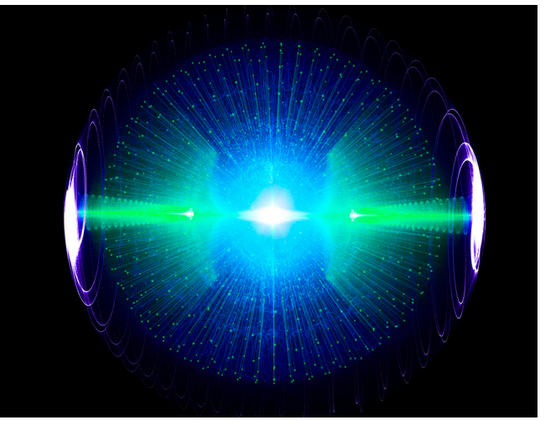
A Generative model references its dataset to concoct coherent language or images, which is part of what has startled and enticed early users. **With natural language programs, while the phrasing and grammar may be convincing, the substance may well be partially to entirely inaccurate, or sometime, when representing a statement of validity, false.** One of the risks with this kind of natural language application is that it can "hallucinate" an inaccurate output in complete confidence. It can even invent references and sources that are non-existent. The model would be forgiven as its function is to generate digital artifacts that look like human artifacts. Yet, **coherent data and valid data are not necessarily the same,** leaving end users of large language models to contend with whether an eloquent output is factually valuable at all.

There is also the risk of inherent bias within the models, owing to the data on which they are trained. No single company can create and curate all of the training data needed for a Generative AI model because the necessary data is so expansive and voluminous, measured in tens of terabytes. Another approach then is to train the model using publicly available data, which injects the risk of latent bias and therefore the potential for bias in the AI outputs.

A fundamental risk is that **users may place complete confidence in erroneous or biased outputs and make decisions and take actions based on a falsehood.** One way to help **mitigate this risk is through AI governance,** and many of the leading practices associated with other kinds of AI also apply to generative models: workforce upskilling, waypoints for decision making across the AI lifecycle, structured oversight, ubiquitous documentation, and the many other activities that promote Trustworthy AI™.

2| The matter of attribution

Generative AI outputs align with the original training data, and that information came from the real world, where things like attribution and copyright are important and legally upheld. Data sets can include information from online encyclopedias, digitized books, and customer reviews, as well as curated data sets. **Even if a model does cite accurate source information, it may still present outputs that obscure attribution or even tread across lines of plagiarism and copyright and trademark violations.**



How do we contend with attribution when a tool is designed to mimic human creativity by parroting back something drawn from the data it computes? If a large language model outputs plagiarized content and the enterprise uses that in their operations, **a human is accountable when the plagiarism is discovered, not the Generative AI model.** Recognizing the potential for harm, organizations may implement checks and assessments to help ensure attribution is appropriately given. **Yet, if human fact-checking of AI attribution becomes a laborious process, how much productivity can the enterprise actually gain by using Generative AI?**

Finding the balance between trust in attribution and human oversight will be an ongoing challenge, with significant legal and brand implications for the enterprise.

3| Real transparency and broad user explainability

End users can include people who have limited understanding of AI generally, much less the complicated workings of large language models. **The lack of a technical understanding of Generative AI does not absolve the organization from focusing on transparency and explainability.** If anything, it makes it that much more important.

Today's Generative AI models often come with a disclaimer that the outputs may be inaccurate. That may seem like transparency, but **the reality is many end users do not read the terms and conditions,** they do not understand how the technology works, and because of those factors, the large language model's explainability suffers. To participate in risk management and ethical decision making, **users should have accessible, non-technical explanations** of Generative AI, its limits and capabilities, and the risks it creates.

Enterprise-wide AI literacy and risk awareness is becoming an essential aspect of any company's day-to-day operations. This is perhaps even more important with Generative AI. **Business users should have a real understanding of Generative AI because it is the end user** (and not necessarily the AI engineers and data scientists) **who contends with the risks and the consequences of trusting a tool,** regardless of whether they should. To promote the necessary AI understanding, CIOs and business leaders may look to existing workforce training and learning sessions, explanatory presentations to business users, and fostering an enterprise culture of continuous learning.

Accountability on the road ahead

Even as Generative AI becomes better able to mimic human creativity, **we should remember and carefully consider the human side of this equation.** Everyone will be affected by Generative AI in one way or another, from outsourced labor to layoffs, changing professional roles, and even potentially legal issues. Generative AI will have real impact, and because **an AI model has no autonomy or intent, it cannot be held accountable in any meaningful sense.**

At scale, the possibility of transparency with Generative AI becomes elusive and "keeping the human in the loop" becomes a growing problem. It is also unclear at this point the degree of consequences that may result from mass adoption of Generative AI, such as the proliferation of fake facts to the detriment of objective and complete truth. These **challenges are unlikely to hinder Generative AI's adoption.**

No matter how powerful it becomes, **we still need the analysis, scrutiny, context awareness, and the humanity of people at the center of our AI endeavors.**

This AI era is the Age of With™, where humans work with machines to achieve something neither could do independently. Now is the time to derive viable methods of accountability, trust, and ethics, linking the Generative AI product and its outcomes with its creator, the enterprise.

Deloitte.