How your organization can have confidence in the opportunities AI brings

by **Beatriz Sanz Sáiz**

Show resources +

5 minute read
15 Jan 2024

Related topics

Technology

AI

As interest in ethical AI explodes, the debate is shifting away from “trust” and towards “confidence”, helping unlock valuable use cases.

In brief

- Large language models (LLMs) are probabilistic rather than deterministic, meaning they carry uncertainty (e.g. hallucinations) with a range of possible outputs.
- A high degree of uncertainty is often intolerable, for example in the healthcare and financial services sectors which combined represent about 40% of GDP.
- There is a clear need for robust governance and stringent risk management to unlock the most valuable use cases of generative AI.

This article first appeared on **LinkedIn**.

In the 1940s, a small number of scholars from a range of fields including psychology, mathematics, engineering, economics, and political science began to explore how to create an artificial brain. At the same time, philosophers, anthropologists, and scientists including Isaac Asimov began to study the ethical implications of doing so. Today, public interest in ethical Artificial Intelligence (AI) is

exploding, sparked by the astounding capabilities exhibited by large language models (LLMs) and generative AI (GenAI) techniques.

According to internet search trends, worldwide search interest in the closely related terms of ethical AI, responsible AI, and AI safety is all at record highs and still rising. However, one related search term is lagging the others in terms of public interest: trusted AI. Why is trust in AI as an ethical framing falling out of favor in the mind of the market.

Building confidence: the next evolution of ethical AI

The EY organization has a long history of working with AI, and it recently announced **EY.ai**, a unifying platform that brings together human capabilities and AI to help clients transform their businesses through confident and responsible adoption of AI. EY.ai leverages leading-edge EY technology platforms and AI capabilities, with deep experience in strategy, transactions, transformation, risk, assurance and tax, all augmented by a robust AI ecosystem. However, as the technology matures and the demands of the market evolve, so do our language, methodologies, and approach to managing the myriad of risks emanating from deploying AI.

As a result of this sustained involvement with EY clients and with AI technologies, we are shifting our focus away from building trust and toward building confidence. This new focus is better aligned to the current and emerging risks raised by leading AI techniques, and clearly better aligned to widespread concerns relating to the ethics of AI. The motivation for this change is quite elemental; LLMs are probabilistic, rather than deterministic, and therefore they bring new risks that we must manage.

In a probabilistic environment, the same prompt or input can generate different outputs, and even seemingly minor perturbations in the prompt such as an extra comma, apostrophe, or alteration in spelling that a human reader would ignore can push the outputs of the model into a completely different space, and as such produce unwanted or unexpected model outcomes (aka “hallucinations”). These new models are also vulnerable to new attack strategies, such as prompt injection attacks where specially designed prompts push the model outside of established safety boundaries. These are additional risks beyond those presented by earlier AI techniques.

How EY can help

Artificial Intelligence Consulting Services

Our Consulting approach to the adoption of AI and intelligent...

Read more

Such variability can’t be tolerated in highly sensitive contexts, like hospitals or clinics for treatments and diagnostics or pharmaceutical processes such as determining doses. With a high degree of model uncertainty, LLMs need to be tightly monitored and controlled in high-risk sectors, applications, or in the automation of business-critical functions. For example, imagine an AI Copilot deployed into emergency rooms to help doctors and nurses triage ailing patients, which to be clear, such products are already in development. A high degree of model uncertainty has the potential to do irreversible harm in such an application.

In turn, this means the incredible technology which has been compared to fire, electricity, and the iPhone in terms of socioeconomic impacts by major technology leaders can’t yet be fully leveraged in key sectors like financial services or healthcare, which together represent around 40% of GDP. As such, many of the most valuable use cases in the economy remain unattainable until strong governance can ensure safety,

robustness, and ethicality. Hence there is a clear and urgent need for rigorous controls to unlock the transformative value of GenAI.

What is required to help ensure safety and robustness are techniques to help us understand the reliability and accuracy of the model prediction, and to measure or predict the degree of uncertainty in its outputs. Techniques are also emerging to further constrain the model outputs sufficiently so that, for example medical experts can be confident that the model is accurate, and that the algorithmic guidance won't do any harm to the patient.

While this is not a solved problem technically, one example is "fine tuning" where additional text is provided to strengthen the model's expertise in a particular expert domain. Another is "reinforcement learning" where humans in the loop provide the model with additional guidance on uncertain edge cases. Next, we can increase the "context window" to provide a source of "ground truth" for the model to reference check, and finally there are "adversarial" approaches where a duelling model essentially reviews and fact checks the outputs of the first model to defined boundaries, such as the language in the United Nations (UN) charter on human rights.

But none of these technical options in isolation or combined are adequate for enterprise adoption at scale. Enterprises need a solution-level approach that encompasses the full model life cycle, from data collection and engineering, to model training and validation, to deployment and real time risk monitoring. Only such a holistic view can give business leaders confidence that the risks are manageable, and that adopting GenAI will be value accretive.

In statistics, the concept of a "confidence interval" is a simple way to describe the degree of uncertainty for a given parameter, providing a range of possible estimates of something unknown. This is a much more appropriate framing for models producing probabilistic outputs.

The growing public debate about ethical AI is therefore more about safety and responsibility. We believe it is possible to ensure AI is safe,

and we believe it is the responsibility of businesses who deploy AI to do so. EY teams are here to help you invest in AI with confidence, unlocking the most value generating use cases to drive exponential growth. Our people-centric approach to AI helps enable the technology to augment your talent, driving efficiency and productivity gains across business functions. And world-leading multi-disciplinary professionals in risk, strategy, technology and transformation help to ensure adoption is aligned with your organization's purpose, culture, values, and key stakeholders so that AI drives positive human impact.

The views reflected in this article are the views of the author and do not necessarily reflect the views of the global EY organization or its member firms.

Summary

A holistic, end-to-end approach to AI risk management is required through the model life-cycle to instill the confidence required for enterprise adoption of GenAI.

About this article

Beatriz Sanz Sáiz
EY Global Consulting Data and AI Leader

Game-changer, thought leader in analytics and customer centricity. Named a Top Talent Executive Under 40 by AMROP, IESE and Royal House. Mom, film lover, champion for women in tech and the workforce.

Related topics

Technology AI

Related articles

How to confidently use AI to create value

12 Jan 2024 | **Beatriz Sanz Sáiz**

G7 AI Principles and Code of Conduct

01 Dec 2023 | **EY Global**

Building a better working world

Connect with us

Our locations

My EY

Site map

Legal and privacy

company limited by guarantee, does not provide services to clients.