Navigating Trust and Safety in the World of Generative AI

By Theodoros Evgeniou, INSEAD; Jeff Dunn and Alice Hunsberger, trust & safety executives

The new generation of artificial intelligence can help defend against online harms - if we can effectively manage the risks.

Keeping up with technological innovations and the debates surrounding their influence on our lives is proving to be extremely challenging for citizens, executives, regulators and even tech experts. Generative AI has ushered in a new era where the creation and dissemination of nearly infinite content has become a tangible reality. Tools like large language models (LLMs), such as GPT-4, and text-to-image models, such as Stable Diffusion, have sparked global discussions from Washington to Brussels and Beijing.

As regulatory bodies race to catch up, critical questions arise concerning the implications for online platforms and, more importantly, trust and safety on the internet. These AI tools may lead to an increase in illegal or harmful content or manipulation-at-scale, potentially impacting our decisions about health, finances, the way we vote in elections or even our own narratives and identity. At the same time, such powerful technologies present significant opportunities to improve our digital world.
It is critical to emphasise that it is not all about an impending AI apocalypse. While *that* is always a possibility – and entirely up to us to avoid – we should be motivated by how we can leverage AI technologies to positively impact our online and offline lives. These tools can be used as weapons for information warfare, or they can be used to defend against online harms originating from both AI and human sources.

Both [Google](https://knowledge.insead.edu) and [Microsoft](https://knowledge.insead.edu) have started utilising generative AI to “supercharge security” and better equip security professionals to detect and respond to new threats. Larger online platforms are already using AI tools to detect whether certain content is generated by AI and identify potentially illegal or harmful content. The new generation of AI can provide even more powerful tools to detect harmful behaviours online, including cyber bullying or grooming of children, the promotion of illegal products or malicious actions by users.

**The good and the ugly**

In addition to reactive protection, generative AI tools can be used for proactive education. One example is tailoring user prompting and policy communications to individuals, so that when they run afoul of a particular platform’s policy or act in a borderline harmful manner, AI tools can step in to promote higher quality behaviour. By regularly guiding and supporting users, AI tools can help everyone better understand and adopt best practices.

For online content moderators responsible for reviewing user-generated content, precision and recall are key. Generative AI can help moderators quickly scan and summarise content such as relevant news events. It can also provide links to related policy or training documents to upskill moderators and make them more efficient. Used responsibly, tools such as ChatGPT or Google’s Bard can also help creators ensure content is aligned with a particular platform’s policies or written in a helpful, inclusive and informative manner.

However, there are various factors that trust & safety policy professionals need to consider before relying on generative AI tools for their daily tasks. Take development of online platform policies for example. Crafting an effective, robust and accessible set of policies typically takes years, involving many consultations with experts, regulators and lawyers. As of now, tasking a generative AI tool with this nuanced work is dangerous or, at best,
imprecise. While these tools can improve the productivity of policy professionals, the extent to which generative AI can be considered safe and reliable for creating and updating policies and other legal documentation remains to be seen.

It is wise to remain cautious and consider the massive volume of content that generative AI can flood the internet with – making content moderation more challenging and costly – as well as the potential harm such content can cause at scale. For example, one of the earliest observed behaviours of large language models is their tendency to “hallucinate” by creating content that neither exists in the data used for their training nor factually true. As hallucinated content spreads, it may be used to train more LLMs. This would lead to the end of the internet as we know it.

To avoid this disaster, there is a relatively simple solution: Humans must be looped into the development of policy, moderation decisions and other crucial trust & safety workflows.

Another problem with LLM-generated content is obfuscation of the original information sources. This differs from traditional online searches where users can evaluate reliability by assessing the content provider or user reviews. Substantial political and social risks arise when users are unable to differentiate between genuine and manipulated content. China, for one, is already regulating the generation and dissemination of AI-generated fake videos, or deep fakes.

Managing the risks instead of imposing bans

The rise of generative AI prompted a wave of discussions about whether technological progress should be put on hold, with thousands signing a letter to this purpose. But while a pause may provide short-term “relief” that we are not hurtling towards some unpredictable AI apocalypse, it is not a satisfactory or even practical long-term solution, especially given the competition between companies and countries. Instead, we need to concentrate efforts on ensuring online trust and safety is not negatively impacted by these technologies.

First, while technologies may be new, the risk management practices and principles employed do not necessarily have to be. Trust & safety teams have been creating and enforcing policy around misleading and deceptive online content for decades and are uniquely prepared to tackle these new
challenges. Common practices for managing other risks, such as cybersecurity, can be leveraged to ensure trust and safety in the world of generative AI.

For instance, OpenAI hired trust & safety experts for “red teaming” exercises prior to the release of ChatGPT. In red teaming, experts challenge a new product in the same way malicious actors would. By exposing the risks and vulnerabilities early on, red teamers contribute to the development of effective strategies and measures to minimise those risks. OpenAI’s now-famous “As a large language model trained by OpenAI, I cannot...” response to potentially dangerous prompts is a direct result of red team efforts.

The skill and creativity needed to be a successful red team member is a burgeoning industry in itself. AI security firm Lakera created “Gandalf”, an AI game to model the problem of prompt injection attacks, where malicious actors inject harmful content into prompts provided to an LLM. To win the game you need to get the Gandalf chatbot to reveal a password seven times. By crowdsourcing “red teaming”, LLMs can be improved to resist prompt injections and other harmful vectors of attack.

Second, guidelines and best practices for how to use these new technologies need to be developed and shared widely. Alongside regulatory efforts, the trust & safety industry is collaborating to develop solutions that can be used by all platforms, ensuring users’ safety no matter where they roam online. The Trust & Safety Hackathon was created so industry professionals can share knowledge and identify such solutions. For example, the industry practice of hash-sharing – sharing cryptographic hashes so companies can quickly identify and remove illegal digital content – has led to a dramatic decrease in child sexual abuse material on platforms.

Third, there will be an increased need to assess the quality of new AI tools, especially as many more versions are being built using “fine tuning” or reinforcement learning from human feedback. A lot can be gleaned from decades of research on evaluating “traditional” AI systems. One common approach is to use statistical metrics such as false positive or negative rates of AI classifiers to measure how accurate these systems are in their predictions. However, assessing generative AI systems may prove more challenging as the quality of their output should not only be measured in terms of accuracy, but also in terms of how harmful it can be.
Measuring harm is difficult as it depends on culture, interpretation and context, among other factors. Similarly, challenges arise when it comes to evaluating the quality of AI tools that determine if content is harmful or not, such as tools that detect illegal products in images or videos. Ironically, LLMs and generative AI can be valuable in evaluating the effectiveness of other AI detection tools and even in managing risks associated with LLMs. It may be that we need more powerful AI in order to manage the risks AI poses.

Finally, after more than a quarter of a century since the dawn of the commercial internet, we need to double down our efforts to increase awareness around online trust and safety. Investments in education around disinformation and scams will help protect individuals from being deceived by AI-generated content that is presented as genuine. The intelligence and analysis provided by trust & safety teams are essential for developing systems that effectively utilise AI to facilitate more authentic connections among individuals, rather than diminishing them.

As our lives have gradually moved largely online, and AI is adopted across industries and an ever-widening range of products, ensuring our digital world is safe and beneficial is becoming increasingly challenging and urgent. Online platforms have already spent many years on their online trust and safety practices, processes and tools. Typically, this work has been invisible, but now is the time for these learnings and experts to take centre stage. We all must work together to chart humanity's path forward as we live alongside AI, rather than being overshadowed by it.

*Jeff Dunn and Alice Hunsberger are trust & safety executives for large online platforms.*

*The lead image was created by generative artificial intelligence program Midjourney using the following prompts: hundreds of computer screens and people in a dark room, minimalistic shapes and cinematic.*

**Find article at**

---

**About the author(s)**
Theodoros Evgeniou is a Professor of Decision Sciences and Technology Management at INSEAD. He has been working on machine learning and AI for over 25 years.

About the series

AI: Disruption and Adaptation

Delve deeper into developments in artificial intelligence, especially the disruptions across value chains. This series examines AI’s impact on a range of sectors, including business consulting, education and the media. It also sizes up the regulatory and ethical questions tied to this game-changing technology.