INTERDISCIPLINARY RESEARCH
AND PROBLEM SOLVING INITIATIVE

ARTIFICIAL INTELLIGENCE, ETHICS,
AND EQUITY RESEARCH GROUP

WORKING PAPER
MAY 2025

Trust and Trustworthiness in Generative Artificial Intelligence

# Why Interacting with GenAI Feels So Safe Yet Remains So Risky

John Carson
Dien Luong
Colleen Seifert

RACKHAM
GRADUATE SCHOOL
UNIVERSITY OF MICHIGAN

# Acknowledgments

This working paper is a research product of the Artificial Intelligence, Ethics, and Equity (AIEE) Research Group. Rackham Graduate School constituted the AIEE Research Group in the fall of 2023, inviting faculty from across the University of Michigan (including Computer Science and Engineering, Information, Business, Public Policy, History, Psychology, and Philosophy) to grapple with emerging societal consequences of artificial intelligence catalyzed by the launch of ChatGPT. The guiding principle for AIEE is to foreground the human in our efforts to understand the implications of AI for society and to envision a framework for ethical AI development, research, and application. This working paper builds on and includes ideas developed through the collective, interdisciplinary discussions of the AIEE Research Group on the theme of Trust and Trustworthiness in Generative AI that took place during its 2024-25 monthly seminar. Not all members of the research group are named as authors, but their perspectives are integral to the conceptualization of the paper.

Members of the AIEE Research Group, 2024-25: John Carson, Joyce Chai, Rita Chin, H.V. Jagadish, Silvia Lindtner, Dien Luong, Nigel Melville, Rada Mihalcea, Joan Nwatu, Shobita Parthasarathy, Peter Railton, and Colleen Seifert.

# I. Authors

**Colleen Seifert** is an Arthur F. Thurnau Professor of Psychology who studies the mind through empirical studies using cognitive science methods. Her research crosses disciplinary boundaries to investigate complex cognition in design, engineering, medicine, AI, and education. Her graduate study in AI and Cognitive Science at Yale University led to her ongoing research on cognitive strategies in natural and artificial intelligence. Her perspective in this work is informed by observations of equity dimensions in technology research and concerns about human needs and rights in sociotechnical contexts.

**John Carson** is an Associate Professor of History at the University of Michigan who studies the entanglements of science/technology/medicine with culture, politics, and society. He focuses particularly on the human sciences in the Euro-American context. After doing an undergraduate degree in philosophy and spending three years teaching in a school for dyslexic children, he did a Ph.D. in intellectual/cultural history and history of science. Much of his work seeks to understand how inequality gets rationalized within liberal democracies. He approaches AI seeking to deflate technological determinist claims about its transformative power by understanding it as a socio-technical artifact whose meanings, powers, and functionalities are embedded in particular cultural, institutional, political, and economic contexts.

**Dien Luong** is a PhD candidate at the University of Michigan studying the complex interplay between geopolitics, social media, and digital governance, with a particular emphasis on Southeast Asia's evolving digital landscape. His research critically examines how online platforms influence public trust, shape political legitimacy, and mediate power relations between Big Tech companies and governments. Having observed firsthand how digital media can simultaneously empower communities and enable state surveillance, his approach in this paper is informed by a cautious balance—recognizing both AI's potential for enhancing democratic engagement and its risks in exacerbating power asymmetries and eroding public trust.

# II. Abstract

Generative AI (genAI) systems are now ubiquitous in everyday activities such as social media use, internet searches, and customer service. How and why might humans place their trust in generative AI systems? What is required to make independent assessments about when genAI is trustworthy? And how can skepticism about genAI help to identify its blind spots, privileged perspectives, and misapplied information? We offer an interdisciplinary approach to understanding genAI contributions as different from those of both human and other artificial intelligences.

Through an analysis of research literature and genAI documentation, we identify the need to approach genAI with "hopeful skepticism" rather than trust. GenAI offers a breakthrough technology for natural language generation, and its ease of use and applicability for diverse purposes are cause for optimism. However, genAI is already clearly problematic in its current form. Understanding when to trust genAI requires considering complex interactions that arise from diverse and sometimes competing interests. These stakeholders include designers and

developers, data scientists and machine learning expert system engineers, project and product managers vetting genAI releases, tech companies producing and selling genAI products, businesses employing genAI applications, individuals or groups or institutions using genAI, and those subjected to genAI determinations. How can regulators, auditors, adopters, users, and subjects of genAI technologies become informed about the value, ethics, and safety of these models to determine trustworthiness?

In this paper, we address the urgent problems of trust in genAI technology and use. First, we identify what is currently known from genAI producers about how genAI technology works and the information provided about its trustworthiness. Next, we investigate what is needed to assess the trustworthiness of genAI expertise by considering some of the factors typically employed when people rely on (trust) human experts. Finally, we identify the importance of context of use, sociocultural perspectives, and biases in the trustworthiness of genAI systems. Our analysis offers recommendations to begin addressing ethical and equity concerns in the use of genAI systems.

# III. Introduction

> *Imagine learning that a new consumer product has just been released. You now experience its recurrent consequences. But you have never seen the new product yourself. You don't know when you are using it. You don't know what it does. You don't know what it knows about you. You don't know who pays for it or benefits from it. No one is accountable for it. You don't even know how it works. In fact, no one really knows how it works.*

People would be unlikely to trust such a product were it a food, drink, health practice, or financial investment. a health practice, or a financial investment. Yet the product described—generative artificial intelligence (genAI)—is already virtually ubiquitous in many parts of the world through its incorporation into social media, internet search engines, educational technology, shopping, and credit transactions, and workplace communications. By law or regulation, many consumer products are required to disclose risks and benefits of use. Currently, AI companies provide minimal information about their products, and while the European Union and other governments have initiated regulation, there are currently no laws or regulations in the U.S.

## What Is Generative AI?

Generative AI (genAI) models are a new type of artificial intelligence that generates new content, such as text, images, audio, and video, using predictive patterns learned from existing data. GenAI models represent a transformational advance in the capacity to respond to natural language prompts by generating novel output instantly. When OpenAI released the text-based *ChatGPT* (OpenAI, n.d.) in November, 2022, it became the fastest-growing consumer application in history, reaching 100 million users in just two months (Hu, 2023). GenAI models have become common to address needs for customer support, content creation,

education, software development, and media marketing – that is, anywhere that the generation of new, original content or data based on existing patterns is helpful.

Also called large language models (LLMs), genAI models "learn" to generate text by identifying word patterns across texts in their datasets. For example, *ChatGPT* (OpenAI, n.d.) used existing data in the form of millions of pages of human-written text scraped from the internet to analyze the relationships between words. *ChatGPT* identifies patterns in texts where words appear together often because they are related. For example, the word "paper" is more likely to appear near "pen" than "shoe." By analyzing how often and which nearby words appear across a mass of texts, *ChatGPT* is able to predict a word likely to follow a string of text.

After training is completed, *ChatGPT* is ready to generate new texts. Starting with words in a user's prompt (or question), a genAI generates a phrase, sentence, paragraph, or whole conversation that echoes patterns derived from its training data. When generating a response, it is not retrieving information from any one of its training texts. Instead, it uses a mathematical summary of its massive number of training texts to predict a connection from a current word to a new one. **Its process is *probabilistic* rather than deterministic**. In other words, based on a given input, it does not predict the same next word every time. It may follow *"The sun rises in the ..."* with *east*, but at other times with *west*, *morning*, or *bread*. A small amount of random variation is added to the computation each time to make its generated text more diverse, like the natural variations in human language (Crawford &. Paglen, 2019).

Based on a user's text prompt, *ChatGPT* can generate a summary, write an essay, translate a passage, tell a story, or answer questions. Other genAI models produce images, videos, music, and even computer code. The great excitement surrounding genAI arises from its ability to generate *novel* output while interacting with users in natural language.

## What Is the GenAI Problem?

While GenAI has already proven to be a powerful and effective tool, it's crucial to understand its risks and limitations. In particular, **concerns have been raised about the trustworthiness of genAI models as sources of information** (Augenstein et al., 2024). OpenAI, the company that produces *ChatGPT*, has provided summary measures of factual evaluation for its models (Figure 1). While improving, accuracy remains below 80% in all categories (OpenAI, 2023). An independent study comparing chatbots' generated citations of scientific papers found 30-90% contained errors, including completely fabricated sources (Chelli et al., 2024).

**Because genAI training processes don't preserve access to sources, there is no way to discern which responses are errors, leaving verification to users.**

GenAI can produce errors or "hallucinations" (Malek et al., 2024; Huang et al., 2023) for many reasons, including insufficient or inaccurate training data, misunderstanding of queries, and compression of data during training. One form perhaps unique to genAI arises from random variation in its probabilistic generation process. At times, it generates new combinations of words that never appeared in any of its training data (Ji et al., 2023). For example, *Google* Bard stated the James Webb Space Telescope, launched in 2021, "took the very first pictures of a planet outside of our own solar system;" however, that happened in 2004 (Milmo, 2023). While so successful in producing original, natural-sounding language, genAI's

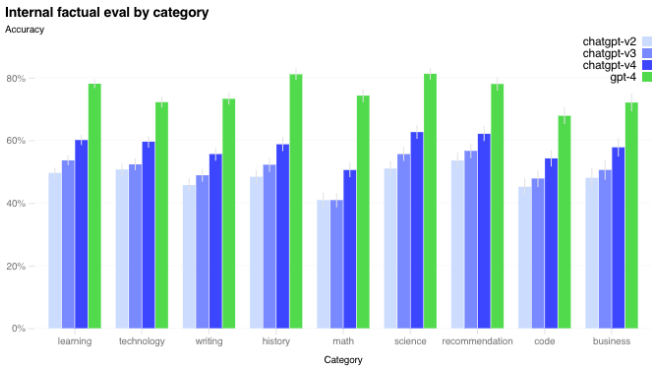


**FIGURE 1**. *GPT-4 (green) shows gains in factual evaluations compared to earlier ChatGPT versions. Accuracy falls below human performance (100%) in all nine categories (OpenAI, 2023).*

probabilistic processes also produce hallucinations that can be reduced but not eliminated (Banerjee et al., 2024).

GenAI is now applied in domains where accuracy is critical; for example, 30% of chatbot interactions during the pandemic asked about "COVID-19" (Chin et al., 2023). Can genAI be trusted for medical advice? Abbass (2019) summarizes the literature on trust and concludes, "Trust exposes a person to vulnerability." This raises the question, how and when must people be especially careful about extending their trust to genAI systems? **What is needed from the producers and deployers (corporations, institutions, governments) of genAI systems to ensure that individuals, organizations, and governments using genAI are relying on trustworthy information?**

Another way to address the question of trust in the context of genAI is to examine how people decide someone is trustworthy. Is this thinking extensible to genAI? In what contexts must people be especially careful about genAI's blind spots, privileged perspectives, and misinformation?

Techno-solutionism ("technology will solve our problems") and techno-optimism ("tomorrow will be better through technological innovation") surround much of the public discourse about genAI. (Roose, 2025; Strain, 2025; Yang et al., 2025) We argue that it is important to counter these prevailing narratives.

**Like any tool, genAI does some tasks better than others. A common misperception is that it generates unmediated and unbiased truth. Instead, users need to think critically about its output and maintain a skeptical posture due to its mix of fact and misinformation.**

Rather than relying on excitement and fears, we adopt "hopeful skepticism" (Zaki, 2024) toward genAI claims and recommend vigilance about its biases, limitations, and partial perspectives. We look for ways to understand the qualities of genAI systems to better harness their enormous data analytic capabilities for social good.

In this paper, we address the problem of how and when to trust genAI systems. We offer three broad perspectives on ways to improve understanding of genAI: 1) transparency about human data, sources, and processes; 2) awareness of how trusting AI differs from trusting human experts; and 3) the need to address blind spots, privileged perspectives, and missing, partial, or incorrect information in genAI output. We offer interdisciplinary perspectives on when and to what degree genAI can be trusted and what protections are needed.

# IV. Trustworthiness, Transparency, and the Producers of GenAI Models

Trust is not just a feature of knowledge, but its foundation. The legitimacy of any information source—whether a government, a journalist, or an expert—depends on whether people believe it to be credible. But genAI is forcing a fundamental shift: What happens when claims are generated by opaque models whose probabilistic processes remain hidden?

**GenAI does not simply offer new ways of accessing information. It reshapes the relationship between knowledge, authority, and trust by presenting generated content whose origins and verifiability differ significantly from traditional sources.**

Unlike traditional sources of knowledge—such as academia, journalism, and public archives—that verify facts through explicit and replicable processes, genAI generates content through probabilistic processes without applying clearly defined thresholds for certainty. While probabilistic *reasoning* is not inherently problematic—indeed, disciplines like law and computer science routinely make decisions based on well-defined probability thresholds—genAI does not weigh probabilities given conflicting evidence, but chooses options based on frequency of occurrence. Without transparency about how companies determine and manage the accuracy of genAI outputs, users cannot critically assess AI-generated content's credibility or potential risks.

Transparency has been identified as a critical need to increase the trustworthiness of AI technologies (U.S. Senate, 2023), yet many governments allow AI companies to dictate the terms of disclosure. While firms like OpenAI, Google, and Microsoft promote transparency, they selectively control access to information about their models. At the same time, state actors increasingly influence these dynamics; recent collaborations between political and technological leaders, such as the alliance between Donald Trump and Elon Musk, underscore how governmental power can shape AI transparency to align with other objectives. Moreover, the rise of global AI players like China's DeepSeek highlights selective transparency when extended beyond Western-centric contexts, reflecting a global struggle over control of AI-generated knowledge.

In this section, we examine how genAI trustworthiness is shaped by its producers, and whether greater transparency can provide meaningful accountability. Generative AI's rapid adoption in governance, law, and media has significantly increased its potential to affect public trust, though the implications have yet to take center stage in public debates. While AI is purported to streamline bureaucracy (Horgan, 2022), it also introduces black box decision-making, meaning the internal logic behind decisions is opaque or hidden from stakeholders. This opacity raises critical concerns about verifiability: how can citizens confirm the correctness of AI- generated legal rulings or administrative policies if the underlying reasoning remains inaccessible?

Recent failures in medicine and law illustrate these risks. In 2024, AI-powered chatbots misdiagnosed patients and recommended fabricated treatment, exposing the dangers of unverifiable AI outputs in high-stakes environments (Burke & Schellmann, 2024). The legal hallucination scandal, in which attorneys submitted AI-generated court filings filled with fictitious citations (Merken, 2025) offers a cautionary parallel. Public

messaging—whether by governments, political actors, or institutions—illustrates how genAI can serve as an accelerant as well as originator of distrust. In these contexts, AI tools are not creating misinformation autonomously but amplifying existing practices of political persuasion and "spin". For example, during the 2024 U.S. presidential election, political campaigns employed genAI to tailor advertisements and fundraising appeals (LaChapelle & Tucker, 2023). Similarly, European far-right parties used generative AI to produce propaganda targeting immigrants and opponents (Quinn & Milmo, 2024), while foreign influence operations leveraged AI-generated content to scale disinformation and manipulate public opinion (Swenson & Chan, 2024). In such cases, AI's role accelerated and personalized human-crafted narratives, complicating public assessments of credibility and intent.

By contrast, other AI-driven harms—such as medical misdiagnoses by chatbots or fabricated legal citations—arise directly from its generative processes rather than its use by human propagandists. Distinguishing between these categories of risk is critical: in the first, AI amplifies human motives, and in the second, AI itself becomes the agent of error.

## Trustworthiness as Defined by GenAI Producers

Public skepticism toward genAI (Pew Research Center, 2025; Gallup, 2023) is fueled by Big Tech's selective approach to transparency. Currently, AI firms in the U.S. dictate the boundaries of transparency, offering selective disclosures while monetizing opacity. Free AI models extract user data, while premium "blackout" models charge for privacy.

**These companies simultaneously market AI as an authoritative tool while legally disclaiming responsibility for its failures. The contradiction is structural, not incidental. These firms want to profit from AI's widespread adoption while shielded from liability.**

A closer look at their disclosures reveals this contradiction: Firms highlight broad categories of data sources while concealing specifics. OpenAI states that *ChatGPT* incorporates licensed sources and human-provided inputs but withholds its dataset composition (OpenAI, n.d.). Google emphasizes that *Gemini* does not use personal Gmail data (Google Cloud, n.d.-a), but offers vague references to training on "publicly available code" (Google Cloud, n.d.-b). Microsoft's *Copilot*, similarly, built on publicly accessible repositories (Salva, n.d.), has sparked legal and ethical debates over its potential replication of copyrighted material.

This selective disclosure is not an oversight. It's a structural feature of AI governance that allows companies to control narratives, manage risks and secure profits. Training data defines the boundaries of what a genAI system "knows" and, by extension, whose knowledge it incorporates to influence its responses—effectively determining whose perspectives are legitimized and excluded.

Beyond corporate strategy, legal and institutional pressures also shape these selective disclosures. Proprietary concerns justify keeping datasets hidden, limiting independent audits. Liability concerns are not just a side effect of opacity—they're a reason for it. If AI companies disclosed their training data, they could face lawsuits over copyright infringement, misinformation, and discrimination. Unlike traditional knowledge institutions, where sources for information,

| Feature | *ChatGPT* (OpenAI) | *Gemini* (Google) | *Copilot* (Microsoft) |
|---|---|---|---|
| **Training Data & Sources** | Public internet data; licensed sources; user-generated data; exact datasets undisclosed. | Publicly available code; no use of personal Gmail data; full dataset unknown. | Publicly available code; no private repositories or personal user data; concerns over copyright. |
| **Usage Guidelines & Limitations** | Not for critical decisions; warnings against use in medical, legal, or financial advice. | Limited to enterprise and developer use; requires AI knowledge; not broadly available. | Code may be incorrect, insecure, or biased; requires human review before use. |
| **Transparency Concerns** | No disclosure of exact datasets; corporate framing of transparency. | Training data is partially disclosed, but without full transparency. | Potential replication of copyrighted material; legal & ethical concerns. |

**TABLE 1.** *Key Disclosures and Limitations of Leading GenAI Models. Information compiled from official disclosures by OpenAI (2024), Google (2024), and GitHub Copilot (Microsoft, 2024).*

justifications, and conclusions are identified as parts of accepted scholarly practice or mandated by law or custom, genAI developers currently face no such obligations. This absence of enforceable standards allows AI companies to define transparency selectively, prioritizing corporate interests over public accountability.

**Transparency is framed as a corporate value, but in practice, it serves as reputational risk management rather than a mechanism for accountability.**

A machine learning researcher may need access to model architecture and weight distributions, a policymaker may need documentation on AI's societal impacts, and a general user may simply need to know whether a response is credible. What AI companies choose to disclose reflects whose scrutiny they are most concerned with avoiding. While opacity may serve as a deliberate corporate strategy, it also has the side effect of reinforcing structural inequalities.

The absence of representative training data is not a theoretical issue; it produces real harm. GenAI overwhelmingly relies on datasets that prioritize dominant voices, often erasing marginalized perspectives (Arnett, 2024). This selective knowledge base means that AI systems struggle with non-Western cultural references, reinforce racial and gender biases, and underrepresent marginalized viewpoints. Studies have documented AI-driven hiring systems penalizing candidates based on race and gender proxies (Basu & Alafriz, 2025), and healthcare risk assessments systematically deprioritizing Black patients for critical care (Obermeyer et al., 2019). Without access to dataset composition, users cannot assess how biases shape AI outputs. And without independent audits, AI's warnings about its own limitations are liability disclaimers rather than meaningful accountability.

Despite withholding full transparency, AI companies strategically emphasize the scale and sophistication of their models in marketing materials. OpenAI describes *GPT-4* as a "large multimodal model" capable of processing both image and text inputs, yet it provides no quantitative specifics about its training data (OpenAI, 2023). Google promotes *Gemini* as its "most

capable and general model," designed to be multimodal and optimized for various sizes, but details on its training sources remain sparse (Pichai & Hassabis, 2023). Each iteration in product releases crafts an illusion of progress, emphasizing capability and efficiency while omitting critical details about dataset provenance, selection criteria, or built-in limitations.

Such selective disclosures serve corporate risk management rather than public accountability. AI firms disclose just enough to comply with regulatory pressure while retaining control over how trust is defined. They engineer a version of transparency that offers minimal visibility without accountability. Without enforceable oversight, AI's credibility remains illusory.

## Why GenAI Cannot Guarantee Accuracy or Reliability

Even if AI companies fully disclosed information about their datasets and model parameters, there would be reasons not to fully trust genAI outputs.

**Transparency alone does not resolve the deeper structural limitations of genAI inconsistencies, lack of accountability, and inability to provide verifiable knowledge. These are not just byproducts of corporate secrecy; they are fundamental to how AI functions.**

GenAI does not assess its dataset so that it will retrieve its output only on well-established facts; it generates instead probabilistic plausible-sounding responses based on material in the dataset deemed relevant by algorithms (though some developers are working on retrieval-augmented generation to address this problem). Accuracy is not guaranteed; instead, genAI offers probability-based coherence. In low-stakes applications—such as brainstorming, drafting emails, or casual content generation—this may be inconsequential. But in law, healthcare, and finance, for instance, where factual correctness is paramount, AI's tendency to fabricate information can have serious consequences.

Even when AI performs well, inconsistencies remain. A study on *ChatGPT*'s responses to breast cancer screening queries found that while 88% were appropriate, others contained outdated or incomplete advice, raising concerns about reliability in high-stakes domains (Haver et al., 2023). These failures are not anomalies but symptoms of a structural limitation: genAI is designed to sound convincing, not to be correct.

AI firms issue disclaimers about these flaws, yet their models are increasingly embedded in industries that demand precision. If accuracy cannot be guaranteed, should AI be trusted in domains where getting it wrong has obvious consequences?

In traditional knowledge settings—academia, journalism, legal archives—consistency, rigor and replicability are prized even if these standards are difficult to achieve. In contrast, the fundamentally probabilistic nature of genAI systems' response generation means that the same input can yield different outputs. Users cannot assume an AI-generated response will be the same twice, let alone accurate. Case in point: A study evaluating *ChatGPT*, *Gemini*, and *Copilot* found that responses to identical prompts varied significantly across instances, exposing the fundamental instability of AI-generated content (Zhu et al., 2024). This is not a glitch but a design feature. Unlike traditional AI, which follows explicit rules, genAI constructs responses on-the-fly—its reasoning is not just opaque to users but, at times, to its own developers.

To be sure, **humans also produce contradictory answers, but trust in human expertise often rests on transparent reasoning, professional**

**accountability, and ethical obligations. In contrast, genAI lacks these institutional safeguards, making its contradictions harder to scrutinize, contextualize, or resolve.**

The absence of such safeguards raises deeper questions about why and how we trust. While humans can rely on peer review, ethical codes, and collective oversight to maintain credibility, genAI offers no equivalent mechanisms, leaving users uncertain about when to defer to its outputs or question its reliability.

## Verifiability: Why GenAI Can't Fact-Check

When a journalist cites a study or a historian references an archival document, their sources can be retrieved, examined, and challenged along with their interpretations. AI, in contrast, generates plausible-sounding information without maintaining links back to sources for verification. This absence makes fact-checking AI-generated content much more difficult and renders responses harder to critique.

But the problem runs even deeper: genAI systems are not just knowledge generators, but some are also data extractors.

> **Every user interaction feeds an ecosystem where data is the currency, product, and raw material all at once. This set of functions creates a dangerous feedback loop: the more users rely on AI, the more their inputs shape future responses, making it impossible to disentangle the basis of genAI outputs.**

Rather than asking why genAI is exempt from traditional disclosure norms, the more pressing question is whether users understand this new kind of tool designed to synthesize information without fixed sources. Ensuring public awareness of genAI's sources draws is essential to prevent its use as an unquestioned source of knowledge. This issue extends beyond misinformation to the very nature of AI-generated knowledge. AI firms argue that training data cannot be disclosed for proprietary reasons, yet monetize AI-generated content derived from uncredited public sources. AI firms demand protection for their intellectual property, so should respect that of others. The Copyright Alliance (Kupferschmid, 2024) has warned that opacity prevents users from assessing AI-generated content credibility, increasing the risk of misinformation proliferation.

A direct comparison between *Wikipedia* and genAI illustrates this problem. *Wikipedia* provides edited descriptions compiled across multiple human authors; while imperfect, it maintains a persistent archive of sources and citations, allowing information to be verified, debated, and updated over time. AI-generated outputs, by contrast, do not persist; the same query can yield different responses, and genAI does not provide a stable record of past answers. This raises a critical question: If genAI's probabilistically generated responses, which lack consistent traceability and verifiability, cannot be retrieved, cited, or cross-checked, can they be considered trustworthy—especially in situations where incorrect information or advice could have serious real-world consequences?

## Accountability: Who is Responsible for GenAI Failures?

As AI permeates governance, business, and media, accountability remains unresolved. Who bears responsibility when AI fails—developers, deployers, or users? Microsoft's *Copilot* warns that AI-generated code may be incorrect or biased, placing the burden of determining accuracy onto users. It is, however, deployed in mission-critical applications. Similarly,

OpenAI disclaims responsibility for *ChatGPT*'s financial, medical, or legal outputs, even as AI automation expands into these fields.

This legal loophole benefits AI firms, who can market their tools as powerful decision-making aids while absolving themselves of responsibility. If AI-generated medical advice leads to harm or AI-powered legal research misleads a court, who is accountable? Without clear mechanisms of accountability, trust in AI systems remains fragile, contingent on whether these systems can transparently demonstrate reliability, responsibility, and responsiveness to consequences. Unlike doctors or lawyers, AI is not held to professional accountability standards.

The problem extends beyond disclaimers. AI firms do not embed explicit and unavoidably-placed warnings into outputs, instead burying liability waivers in terms of service that few users read. Without mandatory, built-in disclaimers directly in AI-generated responses, users risk relying on AI for high-stakes decisions without realizing the dangers.

At the same time, legal frameworks remain fragmented. The National Telecommunications and Information Administration (2023) has warned that AI liability is unresolved, making it difficult for individuals to seek legal recourse. AI regulation is largely reactive: Governments act only after harm is exposed through lawsuits, public outcry, or media investigations, rather than addressing risks at the design stage. Even the EU AI Act, often hailed as a gold standard, faces enforcement challenges, raising concerns that AI firms will simply shift operations to less regulated markets (Bakiner, 2023). Meanwhile, the U.S.'s reliance on litigation results in a slow, inconsistent approach that frequently privileges corporate interests over public welfare (Palmer & Ross, 2024). These gaps allow AI's most harmful effects—algorithmic bias in policing, healthcare, and labor—to persist, disproportionately impacting the most vulnerable.

Ultimately, the question is not whether AI should be regulated, but whether genAI should be treated like any other tool subject to standard liability law. An entirely new legal framework may be needed to address the questions of accountability raised by genAI's novel generation methods. At the heart of the matter is *whose interests will regulation serve*? The corporations and governments developing and deploying genAI technology? Or those the genAI systems are deployed to manage and surveille? To build genuine trust in AI, regulatory frameworks should explicitly prioritize accountability to the public rather than catering to powerful commercial stakeholders.

## Gen AI's Trustworthiness Problem: Moving toward Accountability

GenAI is not just a writing tool. In fact, it is most often used as a mechanism for assessing and generating information, though lacking clear legal and ethical accountability. OpenAI provides guidelines on citing AI-generated content but does not clarify whether AI should be credited as a tool, a co-author, or a knowledge source (OpenAI, n.d.). Google's guidance on citing *Gemini*-generated material remains limited (Brandeis University Library, n.d.). Microsoft's *Copilot* similarly evades authorship responsibility, placing the burden of compliance on users while profiting from its use (GitHub, n.d.). Precisely because of its growing role in knowledge production, genAI requires governance that matches outscale.

We recommend the following approach to enhancing accountability:

### 1. *Mandated Disclosure for Improved Transparency*

Transparency cannot be left to corporate privilege. AI firms selectively disclose data sources, training methodologies, and operational mechanisms, shaping public understanding of genAI while shielding their

systems from scrutiny. If human researchers are expected to disclose their sources, limitations, and methodologies to build credibility, why should genAI producers be exempt? A first step toward improving transparency in genAI systems might include:

- A required authorship statement for AI models detailing the team, their contributions, goals, constraints, and biases in the development of the system (as required in science publications).
- Full disclosure of training datasets, including how data is sourced, curated, and updated, as well as a description of the contents of the dataset.
- Clear, explicit statements on whether and how user interactions are recorded, retained, and then used by the genAI system.
- Independent audits to assess biases, reliability, and systemic risks.

### 2. *Establishment of Legal Liability*

AI firms position their models as authoritative tools and profit from their adoption, while absolving themselves of liability when genAI products make mistakes in mission-critical applications. This contradiction demands regulatory intervention:

- AI-generated content in high-risk domains (law, healthcare, finance) must meet legally enforceable accuracy and accountability standards.
- AI companies must bear liability for harm caused by their models, including through misunderstanding of system limitations by end-users.
- AI disclaimers should be unavoidable on access rather than buried in terms of service agreements and should include, at minimum, attempts to characterize the training data set.

Without enforceable accountability, AI's trustworthiness won't be earned; it will be engineered. Left unchecked, AI's legitimacy will remain a corporate illusion, not a public right.

### 3. *Education for Adopters and Consumers about Appropriate Use*

GenAI companies can provide more information about when and how their genAI models produce trustworthy responses. At minimum, each would post an explanation of its design, projected application contexts, preliminary testing, errors documented and training updates, and disclosures of corrections added to genAI. Errors corrected and improvements made to training, as well as warnings about the probabilistic nature of its responses, would help to inform users about how to consider trust (Knowles & Richards, 2021). One promising approach is AI tool labels (like food labels) providing information on developers, funding, testing, privacy protection, liability, and testing for gender, race, ethnicity, age or disability status biases (Knowles & Richards, 2021).

### 4. *Human Rights Protections for GenAI Use*

In 2021, the European Union passed the Artificial Intelligence Act (European Parliament, 2021), which is currently the gold standard for protection of individual rights in the presence of AI systems. Its provisions are already in effect, and it has set the goal of creating regulatory bodies in each country by August 2026. The act, which represents the first legal framework for AI systems, focuses on risk management, transparency, data governance, and ethical AI practices. It requires that all individuals be informed about an AI system's operation at the time of an initial interaction with or exposure to it. Many familiar applications (such as social media) now employ AI, but people are often unaware that AI agents are recording behavior online. In addition, genAI systems have been used in situations where they wouldn't be expected, such as in

courtrooms. The law thus calls for notice to be given when AI systems begin an interaction or observation of an individual. This stops short of requiring permission to do so and leaves the individual responsible for removing themselves from potential interaction with the AI.

The General Data Protection Regulation extends data protections already in place to interactions with AI systems. These protections assert individual rights, including the right to access, rectify, erase, and restrict the processing of personal data. This provision sets a high standard for protecting personal data and applies to organizations anywhere that offer goods or services or monitor the behavior of EU residents.

The law also requires publishing summaries of copyrighted data used for training AI and designing the AI model itself to prevent it from generating illegal content. The protection of intellectual property and citation of sources helps record and make visible the sources of AI output so that people are aware when they come across such content. Disclosure of AI sources provides important protection for human work.

There are currently few competitors to the EU's AI Act. In the U.S., no national regulation of AI is in place. California has produced the most regulation and will begin requiring genAI developers to disclose information about training data beginning in 2026. Other California laws require oversight for genAI use within state agencies and extend existing data privacy protections to genAI. As laws regulating AI have been passed, some tech activists have been reported as stating they will refuse to comply. An effective system for enforcement will also be necessary given diverse tech interests.

# V. What Makes a Human Expert Trustworthy and Does This Extend to GenAI?

Beyond the issues of transparency and accountability on the part of AI companies and governments, it is worth considering how people determine whether to trust more generally and if this approach can be extended to genAI. What methods may encourage people to evaluate when, why, and which gen AI systems are trustworthy?

## Defining Trust

Trust is commonly defined as meaning, "firm belief in the reliability, truth, ability, or strength of someone or something." The study of why people trust has generated a large literature across disciplines. In anthropology, trust is defined as a social construct varying across cultures; as developed and maintained within specific relationships; and as influenced by social structures, power dynamics, and institutional arrangements (Weichselbraun, Galvin, & McKay, 2023). The primary definition of trust in the social sciences is an emphasis on the expectation of an ongoing relationship:

- A "standing decision" in which someone is given the benefit of the doubt (Rahn & Transue, 1998).
- A relationship among people in which the relationships facilitate ongoing interactions that involve risk-taking and uncertainty about future interaction (Resnick, 2011).

- "Your willingness to embrace the advice of a group of strangers because you believe they (a) know the truth, (b) will tell you the truth as they know it; and (c) have your best interest at heart," all of which depend on "(d) who you are, (e) who they are, and (f) what you're talking about." (Neeley, 2013).
- Willingness to open oneself to risk by engaging in a relationship with another party. (Hon & Grunig, 1999).

Distrust, by contrast, is described along two dimensions. First, the lack of credibility – that is, belief that a person or organization is unaccountable, unethical, or does not respect laws or policies; and second, malevolence – that is, belief that a person or organizations is willing to lie to increase profits, deceive members of the public, or take more than is given (Hon & Grunig, 1999). How well do these definitions capture people's decision to trust others?

## How and Why Humans Trust

Human evolution may have selected for psychological processes to guide people in determining trust in social groups. How we receive, screen, and process information, respond physiologically and emotionally, alleviate uncertainty, detect untruth (cheaters), and accrue reputations contribute to the experience of trust. The evolutionary origins of trust likely occurred in physically embodied, kinship focused, and small communities (around 50) of known individuals (Sutcliffe, Dunbar, Binder, & Arrow, 2012). Human trust is relational, so trusting a stranger does not site with economics' *rational actor model*, which assumes people will take advantage of others if they can (Dunning, Fetchenhauer, & Schlösser, 2019). This framework implies an individual should trust someone only to the extent that they can be compelled to follow through with their promise. It also undergirds contract law, with the government serving as enforcer.

In many parts of the world, people today interact with strangers daily and trust they will not be taken advantage of by others. Trust decisions have the emotional signature of a social obligation or mandate because people typically feel good when they do what they "ought to do" in social encounters. Research shows people behave not because they are giving others the benefit of the doubt but because they are giving in to a social norm. People following the norm of respect feel they cannot insult another person by withholding trust, even if they will never encounter that person again. Reciprocal social norms create trust between strangers and scale trust to interactions far beyond personal relationships (Higgins, 1987). Studies summarized in a recent meta-analysis show that the *reputation* of the trustee (the one who is to be trusted) and the shared closeness of the trustor and trustee were the most predictive factors of trustworthiness (Figure 2).

## How GenAI Might Become Trustworthy

Although it is clear that computers are not people, users often apply social rules and behaviors to their interactions with technology, especially when their use is frequent or prolonged (Madhavan & Wiegman, 2007; Schaefer et al., 2016). People tend to interact with computers in ways similar to interacting with people. Computer tools such as automated decision aids are often intentionally designed to emulate human interactions in language and social etiquette (Nass & Moon, 2000). New genAI systems with natural language interfaces have pushed this potential even further by allowing conversations rather than command-line interactions.

**As a result, people may apply the same assessments of trustworthiness in humans to genAI.**

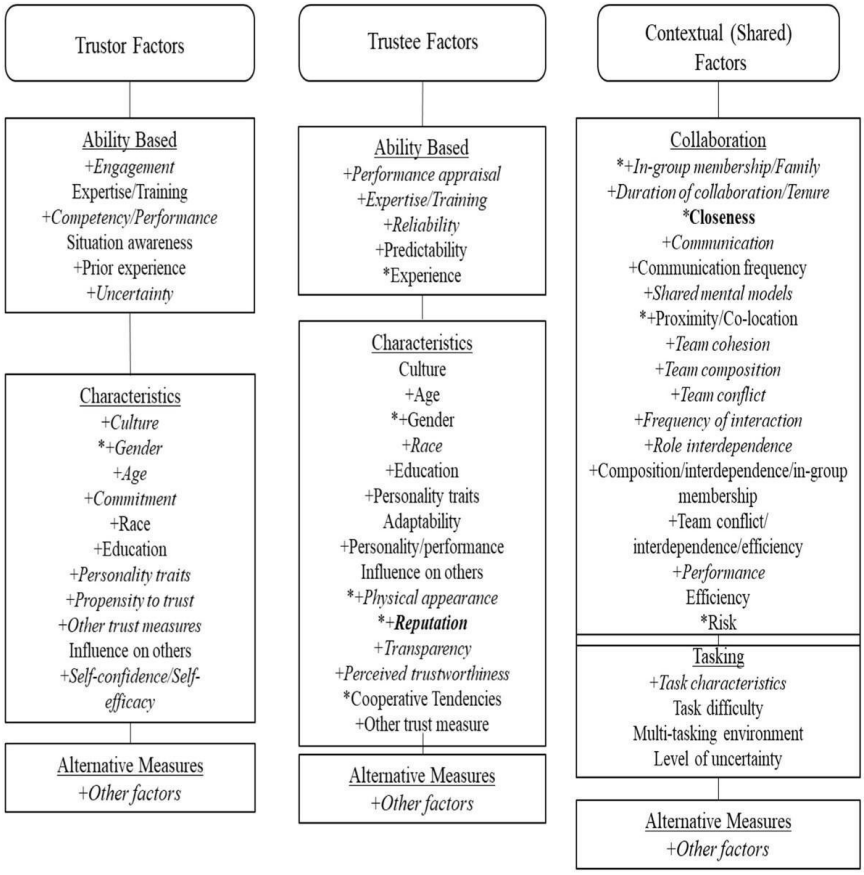


**FIGURE 2.** *Empirically identified factors influencing trust judgments. Terms with a (+) represent correlational findings and a (*) experimental findings and italicized and bolded terms represent statistically significant findings. (Hancock et al., 2023).*

Drawing from the model by Hancock and colleagues (2023) (Figure 1), should people apply criteria for human trust to genAI? The contextual factors people use to assess trust include relationship closeness, task qualities, and teamwork dynamics. These factors raise major challenges in assessing genAI as a trustee with the same qualities. Many qualities are not applicable to genAI, but others could be made apparent with greater transparency about genAI.

Physical appearance and social identities are important in assessing human trustees but are not currently applicable for genAI. Similarly, personality traits are not distinctive (yet) for genAI. People can assess a genAI for qualities such as cooperativeness, adaptability and performance. However, other genAI characteristics – such as expertise and training, reliability, predictability, experience, and transparency – are currently difficult to assess. If more information were disclosed about training and performance across users, assessing these qualities could aid people in evaluating trust in genAI. Further, while reputation is an important quality for trusting in a human, there is currently little information shared across people regarding genAI and its influence on others. Because fewer factors about genAI as

trustee are available for people to consider, it may be challenging for people to judge whether to trust an AI system and may lead to assessments of (dis)trust.

Another source of information about trustworthiness is contextual factors. Feelings of closeness are the most predictive, but it is clear that genAI use can evoke emotional responses leading to trust, and at times inappropriate reliance (Saracini et al, 2025). Less is known about how specific genAI applications may manipulate users to gain trust. Other factors important in trust involve collaboration and task qualities. In addition to co-location and frequent contact, communication, cohesion, and performance lead to increased trust. Experiences of human-genAI teamwork will likely allow people to feel more comfortable trusting genAI.

The evidence about trust in genAI shows most people are wary (Gillespie et al., 2023). People trust AI most in healthcare contexts and least in human resources. While most (85%) recognize AI's benefits, only half believe they outweigh the risks, with cybersecurity rated as the top risk globally. In comparison, people are most confident in trusting universities and defense organizations (79%) and least confident in government and commercial organizations (33%). Given U.S. political events since 2024, concern about trust in public institutions is likely even higher as basic principles of law, ethics, and trust are undermined in current policymaking and governance.

In line with these findings, almost all U.S. respondents (Gillespie et al., 2023) strongly endorsed the principles defined by the European Commission (2019) for trustworthy AI:

(1) lawful - respecting all applicable laws and regulations;

(2) ethical - respecting ethical principles and values;

(3) robust - both from a technical perspective and social environment.

To accomplish these principles, **people want regulation with external, independent oversight**. In the workplace, about half are willing to trust AI to augment work and inform managerial decision-making, but they want humans to retain control. While people indicate openness to using AI, over half feel they don't understand it, including when and how it's used; for example, that AI is incorporated in social media. Those who have a better understanding of AI are more likely to trust it (Gillespie et al., 2023).

In fact, trust in AI may **exceed** assessments of trustworthiness. There are examples of people following an AI's advice even when it contradicted contextual information and their own assessment (Klingbeil, Grützner, & Schreck, 2024). An overriding trust in AI may be due to its presumed "objectivity" as a computational system, and the hype surrounding the rise of genAI may contribute to this.

**To determine whether and how much to trust genAI systems, people must understand them and have access to information to assess when, how, and which genAI to trust to what degree and under which circumstances. Assuming an AI is superior to human expertise without assessing the system's trustworthiness for the task at hand may lead to undesirable outcomes.**

## Transparency as a Proxy for Trustworthiness

With genAI, people need to (1) trust a prediction to act on it, and (2) trust it to behave in reasonable ways if deployed (Riberio, Singh, & Guestrin, 2016).

| *Not currently available* | *Currently available* | *Potentially available* |
|---|---|---|
| Culture | Cooperative tendencies | Expertise/Training |
| Age | Perceived trustworthiness | Reliability |
| Gender | Adaptability | Predictability |
| Race | Performance appraisal | Experience |
| Education | | Reputation |
| Personality traits | | Influence on others |
| Personality performance | | Transparency |
| Physical appearance | | |

**TABLE 2**. *Qualities of a trustee that influence human trust in others have been identified in empirical studies (Figure 1; Hancock et al., 2023).*

To calibrate trust in genAI, people need to understand it. However, many genAI products hold back information that would be helpful. Does genAI have to be transparent for humans to trust it? Ross (2024) argues that requiring transparency to examine ethics and fairness is holding AI to an "epistemic double standard." People appropriately defer to human experts and accept their advice without necessarily understanding it. As Ross (2024) points out, "If I knew what you know, I would not need to trust you." In this sense a genAI "black box" is not different from the human expert "black box." Transparency itself need not be the goal.

Ross (2024) suggests the reason for the demands about AI transparency is its role as a *proxy for trustworthiness* (Ross, 2024). Trust is essential with both human and AI experts *because* of the absence of understanding and explanation. As Microsoft CEO Satya Nadella (2016) stated, "We want not just intelligent machines but intelligible machines. People should understand how the technology sees and analyzes the world." What information about a genAI system is needed to make AI expertise intelligible to humans? Ross (2024) proposes that, "trust and trustworthiness can replace understanding as epistemically and ethically sound grounds for belief."

With human experts, grounds for trust are provided by the existence of social institutions that provide verification. Based on qualitative analysis, we have identified evaluative features laypeople use in assessing whether to trust a human's expertise. These include knowledge and ability, past performance, institutional framework, professional standards, social and community relationships, and explainability. A surprising number of considerations regarding trust in an expert are driven by qualities other than ability.

## Standards of Expertise for Humans and GenAI

### *1. Knowledge and Ability*

Knowledge and ability—including knowledge base, process, and domain coverage—are key attributes of human expertise along with both academic training and field experience. Little information is available about AI baseline capabilities (Jones, 2024); however, AI successes (e.g., competition-level mathematics)

| Standards of Expertise | Features | Qualities | | | Type | |
|---|---|---|---|---|---|---|
| | | | | | Human | genAI |
| **Knowledge and Ability** | *Knowledge base* | Timely | Relevant | Contextualized | ☑ | ☐ |
| | *Process* | Justification | Use of evidence | Analytical voice | ☑ | ☐ |
| | *Domain Data* | Access | Scale | Representativeness | ☑ | ☐ |
| | *Experience* | Varied cases | Varied contexts | Years in Field | ☑ | ☐ |
| **Past Performance** | *Accuracy* | Quality | Measures | Errors | ☑ | ☐ |
| | *Reliability* | Field consensus | Citation | Second opinion | ☑ | ☐ |
| **Institutional Framework** | *Accountability* | Legal | Contractual | Insured | ☑ | ☐ |
| | *Legitimacy* | Credential | License | Awards | ☑ | ☐ |
| | *Standards* | Defined field | Authority | Peer Review | ☑ | ☐ |
| | *Sources* | Attribution | Citation | Insider status | ☑ | ☐ |
| **Professional Standards** | *Independence* | No outside interests | Fees and services | Intellectual property | ☑ | ☐ |
| | *Objectivity* | Independent | Free from bias | Disclosures | ☑ | ☐ |
| | *Ethical Values* | Privacy | Confidentiality | Humility | ☑ | ☐ |
| **Social & Community Relations** | *Social standing* | Reputation | Recommendations | Ratings | ☑ | ☐ |
| | *Personal* | Invests | Reciprocal | History | ☑ | ☑ |
| | *Encounters* | Rapport | Listening | Shared world view | ☑ | ☑ |
| | *Liking* | Similarity | Proximity | Interactions | ☑ | ☑ |
| | *Exclusivity* | Prestige | Customized | Concierge | ☑ | ☐ |
| **Explainability** | *Transparency* | Explanation | Expectations | Uncertainty | ☑ | ☐ |

**TABLE 3**. *Summary of elements in human expertise as defined by professional fields (Ross, 2024) Social institutions create and maintain standards for professional experts in a field. Most of the elements are undefined for genAI systems. While some standards (e.g., Encounters) are defined for individuals, community consensus is not available for genAI (yet). Many elements are potentially definable but currently hidden in genAI (e.g., Past performance) or currently undefined (e.g., Accountability).*

tend to occur in areas with well-defined problems and solutions (e.g., games like chess and GO). GenAI's new probabilistic process is now employed in domains such as hiring decisions, student admissions, therapy, recidivism predictions, micro-loan applications, and business forecasts. These are applications where there is a need for expertise, but also limited evidence that even human experts can routinely produce reliably good outcomes.

One point of comparison for OpenAI's GPT models has been standardized aptitude and achievement tests such as the LSAT law school admissions exam. Comparing genAI performance relative to a large population of people, *GPT-4* scored at the 88th percentile on the LSAT (OpenAI, 2023) and it passed the uniform bar examination (Katz et al., 2024). Ideally, genAI performance would be compared to human experts on the same new problems in a field. One comparison in medical radiology for prostate cancer diagnosis. Because the data is limited to images, human and AI experts can be compared using the same data and judgments. The diagnostic performance of deep learning models was equivalent to health-care professionals, with pooled sensitivity around 87% for both (Xio et al., 2019). However, human experts also offer flexibility, such as identifying whether another

scan would be informative. Empirical studies with direct comparisons to humans are important to demonstrate the adequacy of genAI determinations.

As noted earlier, the sources and nature of training data are considered proprietary and not shared with users or adopters. However, awareness of the kinds of data included, such as specialized (academic publications) or general sources (web scraping), text languages and geographic areas, and scope would be informative for users. Without disclosure, there can be no consumer power to choose among AI tools based on their comparative features.

### *2. Past Performance*

People also determine trust in human experts based on performance, including accuracy and reliability. How often, quickly, and well are users of genAI served? There are currently no efforts to collect feedback from users or to estimate speed, response length, or comparisons across similar queries. Performance measures are rarely supplied to customers even when a genAI company reports improvements (OpenAI, 2023). Despite many decisions made by genAI daily, no information about success is available beyond what occurs in an individual's session. In some fields of expertise, performance outcomes are available from court records, professional societies, or Better Business Bureau complaints. Longevity of practice and number of customers served are other indicators. Puzzlingly, little comparative evidence has been provided by companies offering commercial genAI despite highly competitive marketing (Figure 3).

In one study, texts were provided to genAI models and their responses to queries assessed (Bahak et al., 2023). Evaluation of hallucinations revealed that ChatGPT provides responses to a significant percentage of questions for which no answer is available in the provided context. The "generative" process in genAI creates hallucinations as a natural consequence of its probabilistic processes. While expected to decrease with richer training datasets, hallucinations can never be eliminated (Banerjee, Agarwal, & Singla, 2024). The takeaway for users is that they should not trust or put into public use any information produced by a generative AI model without checking its accuracy, especially in high-stakes domains and professional settings. For example, genAI in automated vehicles learns to recognize road signs from training images. But if a sticker has been placed on a stop sign, its appearance is changed enough that genAI may not recognize it. A stable (non-genAI) algorithm works almost as well when exposed to a small amount of noise in its input, though exactly how much requires testing. Research suggests, however, that it is mathematically impossible to develop universally stable genAI algorithms (Chase, Moran, & Yehudayoff, 2023).

### *3. Institutional Framework*

Institutions define expertise by consensus within a specialized community. Independent experts in a domain regularly converge in opinions to produce a reliable judgment of correctness and a means of resolving disagreement. Technical specializations are established in educational programs and professional organizations to create standards of practice and grant degrees, credentials, or certifications. Some professions (such as medicine) impose approvals, rules, and legal penalties, and may revoke credentials.

How will such institutional structures be created for genAI? Who are the stakeholders defining generative AI? Tech companies view similar experts as competitors and define qualities as proprietary. A network of specialized experts is needed to document performance standards, such as whether genAI use in hiring decisions is well grounded. Government regulation and legal accountability lags far behind development. Regulatory proposals offer general guidelines without defining expertise, such as the

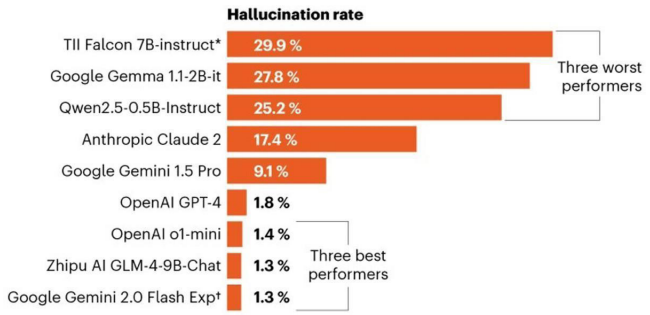


**FIGURE 3.** *The Technology Innovation Institute tested LLMs on their performance on the task of summarizing a provided text. The best- and worst-performing genAI systems are shown, with the well-knowns ChatGPT producing 1.8% and Claude 17.4%.. Size did not appear to impact accuracy, as an OpenAI "mini" system performed better than GPT-4.*

European Commission's 2021 proposal for Regulation on Artificial Intelligence, the UK Institute for Ethical AI and Machine Learning, the Global Partnership on AI (GPAI), and the OECD AI Policy Observatory. Current proposed "credentialing processes" for AI systems focus on technical execution, established by societies such as the Institute of Electrical and Electronics Engineers (IEEE), the International Organization for Standardization (ISO), and the International Electrotechnical Commission (IEC). That leaves genAI users to organize crowdsourced comparisons, but currently few places host shared information about successes and failures with genAI.

### *4. Professional Standards for Conduct*

Another important part of assessing human experts is professional conduct. Experts are hired in an agency role for individual clients to act on their behalf. Codes of conduct typically require them to disclose fees for services and identify themselves as independents or agents with a company. If a financial advisor receives "kickbacks" from their agency for selling specific stocks, the advisor's other interests must be disclosed. Clients may also discuss values with experts to see if they are aligned (e.g., environmental concerns, risk comfort, cooperation versus competition) and assess the standards in the expert's practice, such as privacy, confidentiality, and record keeping. Human experts want clients to return or to recommend them to others, incentivizing their performance.

Key stumbling blocks for trust in genAI include lack of individual relationships with users. Currently some users have genAI track their use across sessions, but others do not, and may use multiple AI tools. Recommender systems and social media follow an individual's genAI use and collect data in later interactions, but this is not disclosed. Without a personal stake in each client, there is no pretense that the AI is "looking out" for individuals, a key feature of trust in human experts.

An important difference between human and genAI expertise is motivation or intention. Human experts are incentivized by pay for performance, but the genAI itself is not rewarded by its success. The outcomes of success include financial, reputational, and intellectual satisfaction for developers and companies, but not for genAI. This dynamic highlights the presence of multiple stakeholders in determining AI behaviors, such as developers, advertisers, companies, and stockholders. Do these interests compete with a user's?

### *5. Social and Community Relations*

A community of laypeople who interact with the same human experts may pool information, formally and informally, to define a reputation through anecdotes about success rates, personal qualities, and client relationships. Patient ratings of health professionals form a separable dimension of expertise defined by laypeople. For example, the investment of the expert in the client (such as their memory for past encounters) can be assessed by other clients' reports. The closeness

of a referral (a family member, friend of a friend, or online source) may influence whether a client selects a given expert.

In other service and product domains, many online ratings sources are offered by client communities and third parties such as (in the U.S.) *Yelp*, *Consumer Reports*, and *Amazon* ratings. At present, competing genAI systems are fighting for dominance in the market, but no shared information from users is providing reputational and relationship incentives for genAI. Further, human experts are expected to meet particular moral and ethical standards, sometimes made explicit by professional societies. AI appears to need interventions to pursue values such as avoiding harm. Some genAI models have parameters encouraging cooperation following a user's lead. For human experts, the possibility of free initial interviews allows a client to assess whether their needs align with a provider; genAI, however, lacks any purposeful inquiry to help it align with a specific user's interests.

### *6. Explainability of Expertise*

In most cases, neither genAI nor human experts can provide a description of their process in arriving at a decision (Liu, 2019). Rudin (2019) argues that "black box" models uncover hidden patterns human observers cannot detect. AI is most useful in areas where its processing capacity surpasses human abilities, resulting in opaqueness. As Riberio and colleagues (2016) write, "... if hundreds or thousands of features significantly contribute to a prediction, it is not reasonable to expect any user (or developer) to comprehend why the prediction was made, even through inspection of individual weights."

What happens when a human expert is asked how they reached a decision? If they remain mute, would they be trusted? A human expert would be expected to describe the basis for their decisions that makes sense to a non-expert. They may include experience over time, statistical outcomes, or examples of similar cases. An expert may claim the reasons for their decision are not knowable or understandable to a layperson, but they would still maintain that it has a well-grounded basis. Their expertise is describable even if their knowledge and process is not.

People are much less likely to trust experts who are unable to communicate with non-experts. It is important to demonstrate to the client that the expert understands the application of their expertise to the client's specific issue. This is where the explainability problem arises in genAI (Ali, 2023). A client needs assurance that the factors considered are relevant to their case. For genAI, the knowledge base and process generate output without any ability to understand or reason about what it is computing. Human expertise appears more versatile and articulable, suggesting genAI expertise may be less valuable. However, genAI's lower cost and greater availability offer other advantages.

**Further, a human expert and genAI differ in their capacity for metacognition, or "thinking about thinking." Awareness and control of cognitive processes includes knowing when a conclusion is uncertain, when information is false, and when information is not known or not knowable. GenAI systems are just beginning to add processes to simulate cognitive awareness about information status. Metacognitive processes create awareness of the status of information as true or untrue, what data is missing, and how knowing it may change the outcome. The ability to reason about the connection between meaning in the database and**

**the status of information generated in responses points to a major difference between human and AI expertise.**

The inability of genAI to assess its generated knowledge poses a barrier to use. A human expert can indicate when they don't know with accuracy, but a genAI is unaware that a hallucination is in error (the AI states it is true and affirms it) through confabulation (combining pieces of information in the training data to create new text not appearing in any of them). Some genAI systems are beginning to use sampling to measure certainty; that is, running repeat queries to determine whether there is consensus among responses. Does a genAI "know what it knows"? Some genAI models can state confidence in the correctness of an answer, where perfect calibration is the same accuracy as stated confidence; however, findings show overconfidence in all genAI tested (Figure 5; Kadavath et al., 2024). Creating a method to determine whether a genAI has produced an accurate answer may be as challenging as devising a correct answer.

A human expert who stated their confidence as over 90%, but whose accuracy was 50%, would give a client pause. GenAI systems may be given more grace because their actual performance is not advertised. Mitigation begins by informing people about the fact that errors occur often. Currently, little information is provided about genAI accuracy by its producers. GenAI apps should caution before each use that they are not suited to answering simple factual queries. In addition, people should be made aware of the multiple causes of errors in genAI. Chatbots are also designed to produce a response that fits the situation, so they tend to agree with a user's errors (Suzgun et al., 2024). Newer models are even more likely to answer with a response, resulting in more mistakes (Zhou et al., 2024). Currently, informing users about genAI accuracy is needed as a "buyer beware" warning before each genAI use.

In sum, establishing institutions to assess the trustworthiness of genAI requires the creation of socially-defined, consensual definitions of expertise, performance, and practice.

**Steps for determining genAI trustworthiness:**

1. Certification of knowledge base, process, domain data coverage, and training data content.
2. Certification of standardized performance accuracy measures for specific genAI models.
3. Creating institutional frameworks for expertise across genAI models and domains.
4. Establishing standards of conduct such as confidentiality for genAI systems.
5. Disclosure of social community responses to genAI performance experiences.
6. Explanation of a genAI's expertise and its application within the context of a query.

# VI. GenAI in Context: Trustworthiness, Bias, and (Mis)Information

Context matters for the trustworthiness of genAI, with different consequences depending on who is impacted by its use, their purpose, social identities, and national/cultural contexts (Jacovi et al., 2021) In general, trustworthiness in a genAI system must be high when making or advising consequential decisions with substantial risk, or what Ross (2024) calls "'*ethically significant*' contexts of use." For example, determining eligibility for social services, or screening job applications, or reviewing criminal records raise the risk of increasing inequity. In other situations, such as planning a holiday trip or checking for grammatical mistakes, the costs of error are relatively low.

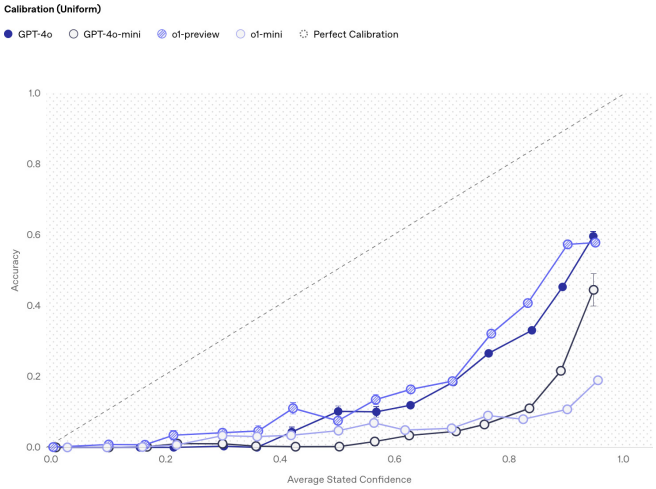


**FIGURE 4.** *Larger genAI models display better calibration of correct performance and confidence; however, all models tested fall well below the line (y=x), meaning that models consistently overstate their confidence in their accuracy (Kavadeth et al., 2024).*

## GenAI and Contexts of Use

Humans regularly use information they know (or suspect) is not fully accurate and trust sources with diverging interests. In principle, genAI should be treated no differently; it provides information that may be useful but demands scrutiny when consequences are significant, including checking other sources if possible. Abbass (2019) argues that a critical determinant of trust is whether the intent of the genAI system aligns with the intent of the user (or individual subjected to its use). However, unlike most tools, intention in genAI systems can be difficult to identify. In practice, the speed and ease of information access, perceived objectivity, and advertised scope make it easy to accord more authority to genAI outputs than other information sources. Thus, it is important to consider how a genAI system's trustworthiness is, in part, a function of context. (See Table 4).

For highly consequential situations where genAI is making most of the important decisions, the ideal is a genAI system trained on a well-curated data set directly relevant to the situation and cultural context of use. The model's "temperature" must be set to produce reliable responses (i.e., not vary significantly when repeated), various testing procedures run under trial scenarios, and some indication of uncertainty ideally provided (Afroogh et al., 2024). Periodic retesting would also be necessary to confirm that the genAI system is working properly or updating of the dataset if it has been in use for a long period of time (Quach, 2020).

One approach to determining the trustworthiness of a genAI system in a given use situation is to ask the following questions about how the genAI system will be employed (Spiegelhalter, 2020):

- What is the need to be addressed? How consequential will decisions be to those affected?
- Is genAI a good tool for the need, given that genAI models are probabilistic and nondeterministic, but also able to summarize and simplify complex information?
- How much authority will the genAI system be granted in addressing the need?
- Are those using or subjected to a genAI system in a position (social, economic, educational, cultural, etc.) to challenge or contribute to its outputs (if desired)?
- Is the intent of the genAI system aligned with the intents and interests of those using or subjected to it?
- Are there ways to supplement use of the genAI system with other sources of information or expertise?
- How catastrophic would failure of the genAI system be for whatever reason?
- What alternatives are available?

Gerd Gigerenzer (2023) suggests that one important way to determine how much to trust an AI's output is to distinguish between two situations: 1) those with well-defined, stable rules with large amounts of available data (such as playing chess) and 2) those with ill-defined, unstable operations with a high degree of uncertainty. While genAI performance is very good with the former, so-called "small world" problems, it continues to struggle with the ambiguities of ill-specified real-world situations with complex interactions.

## Privileged Perspectives and Blind Spots in GenAI

As noted in prior sections, lack of transparency is one of the most important impediments to assessing the trustworthiness of genAI-produced information (Felzmann et al., 2019). Major determinants of output—training dataset, association weights, and generation algorithms—are not provided to users or third parties *by design*. Because genAI models create responses using probabilistic algorithms, they cannot identify the specific sources drawn on for their outputs nor assess truthfulness. (Zhou et al., 2023) They can generate references; however, in most AI models these citations are independent of the model's response sources and may be nonexistent (hallucinations). GenAI producers regard training datasets as proprietary, so they are rarely revealed. Consequently, users have no idea which sources contribute to genAI-produced information. To address the transparency issue, explainable AI (xAI) approaches attempt to add post hoc accounts to explain how responses were generated, though to date with limited success. (Afroogh, 2024; Jacovi et al., 2021; Kamath & Liu, 2021; Molnar, Casalicchio & Bischi, 2020) Some other new genAI models add estimates of uncertainty through comparison with a sample of responses from other genAIs (OpenAI, n.d.).

Critically, genAI models rarely address the issue of biases inherent in their database or processes. One suggestion is to promote regulations requiring that the initial training data for a genAI system be characterized thoroughly to provide some insight into the orientation of the genAI model. Knowing that a genAI was trained on data scraped from English-languages sources ranging from *Facebook* to *Wikipedia* to *GoogleBooks*, for example, might inform users about its intellectual, epistemic, and cultural presumptions and suggest cultures and communities absent from its knowledge base, as shown in prior technologies (Lee & See,

| GenAI Role | Highly consequential decisions | Less consequential decisions |
|---|---|---|
| **Embedded and makes most decisions** | Cars operated in self-driving mode without a human driver | Automatic grammar and spelling checker |
| **External and used to make or justify decisions** | DOGE decisions to eliminate government positions | Chatbot providing music suggestions |
| **High influence with other sources available** | Use of facial recognition AI by police investigators | Students getting homework help and checking their work |
| **One of many sources of information** | Business strategic decision- making such as identifying potential markets | Travel recommendations for a vacation trip |

**Table 4.** *Eight varied contexts for the use of AI raising different issues in the degree of trustworthiness necessary. The need for trustworthiness is highest in the upper left and is least important in the lower right.*

2004). However, the utility of assessing biases through the examination of training data descriptions is open to question. Would users be able and willing to invest in considering what training data may imply about biases in a genAI?

A second approach is to pose the same query to multiple genAI systems and examine differences in responses that emerge. In principle, if the genAI systems were built from very different training datasets, this strategy might illuminate the impact of training data on genAI system output. In practice, however, the major genAI systems currently available (with the exceptions of *Deep Seek* or *Qwen*, produced by Chinese companies) were developed by U.S. companies and drawn from similar readily available English-language digital materials. As a result, genAI may well represent the same broad cultural outlook in every available product.

A third possibility is to turn away from general purpose AI systems to ones built with targeted expertise. Some genAI training datasets have been curated to contain more limited and vetted data sources for special purposes. For example, a genAI for medical use trained on data from medical journals, textbooks, and professional material may provide more accurate answers from a medical science perspective and make more evident the absence of alternative approaches (Eastern, holistic, spiritual). Highly specialized medical AIs using limited datasets such as radiograph images are among the most successful applications of genAI. While such targeted genAI systems would still have blind spots, areas of low information, and potential for misinformation, these gaps and biases might be easier to anticipate.

## (Mis)information and Biases in GenAI

The online world of opinions, beliefs, errors, facts, and intentional disinformation has allowed misinformation to spread widely (Carroll, 2024; Center for Countering Digital Hate, 2024), though some scholars dispute the severity of the issue (Altay, Berriche & Acerbi, 2023; Arechar et al., 2023; Fletcher & Nielsen, 2019). The complex network of information sources, technologies, and audiences, called the "(mis)information ecosystem," does not identify information source and

thus allows anonymous and false sources to promulgate information and misinformation (Wanlas et al., 2025). While the media (newspapers, magazines, radio, television, etc.) have also promulgated information of dubious validity (Zhou et al., 2023), the internet and social media provide the means for individuals and groups to spread ideas to much wider audiences much more quickly than previously possible. As output from genAI enters this ecosystem, what challenges does it pose for distinguishing between information and misinformation, and who decides the criteria?

On the surface, genAI may not fundamentally change the problem of assessing the truth of online images, text, and video even if it adds to or accelerates it in some ways (Carroll, 2024; Chakravorti, 2024; Merchant, 2025; Wanlas et al., 2025; Zhou et al., 2023). GenAI outputs can include specific markers of authenticity, such as citations, specific details, allowances for discrepancies, and analytical voice. GenAI can also avoid some misinformation cues, such as poorly formed arguments and incoherence, making genAI misinformation more difficult to detect (Zhou et al., 2023). Yet, as Altay, Berriche & Acerbi (2023) argue, "people are not passive receptacles of information;" rather, they "domesticate technologies in complex and unexpected ways" (c.f. Fletcher & Nielsen, 2019)

AI-generated claims (even when dubious) enjoy a boost from the tendency to perceive machines as purveyors of objective facts. Human bias is widely known and judged, but computers are associated with unbiased execution of rules. However, genAI poses specific impediments to information accuracy. The problem of hallucination occurs when a genAI model produces inaccurate, invented, or misleading results and presents them as factual. While research shows that rates of hallucination decrease with database size, they cannot be eliminated (Chelli et al., 2024). GenAI images often have added fingers or distorted backgrounds from ambiguities in the images provided during training. For example, asking *ChatGPT* for an image of a clock showing a certain time typically produces one with the hands at ten minutes to two, regardless of the time requested, because that is the most common image used in advertising (Block, 2024). GenAI summaries display problems detecting tone, such as sarcasm, and in foregrounding opinions derived from experts on the issue being queried.

A deeper problem with misinformation derives from biases inherent in the datasets provided to genAI models (Afroogh et al., 2024; Mihalcea et al., 2024). When asked for an image of an 18th- century dwelling, *ChatGPT* returned a Georgian mansion, stately and imposing but certainly not representative of the kinds of habitations in which most people in the 18th century resided. When asked what a typical dwelling looked like, it acknowledged that dwellings "varied by region, social class, and available materials" (Carson, 2025). Nonetheless, *ChatGPT* referenced only types of dwellings found in the American colonies and western Europe, excluding residences of poor people, indigenous groups, and people living in the rest of the world. The output *ChatGPT* provided was not false but reflects a database where images of elite Euro-American homes are much more prevalent. Ethnocentrism can be pervasive in any information source, and genAI systems be trained to suggest more diversity in responses (Milhacea et al., 2024). But the combination of the seeming objectivity of a machine-derived response, source opacity, institutional powers, and interests of AI producers result in covert ethnocentrism in genAI that reinforces existing social biases.

Finally, genAI systems may be particularly liable to plagiarism, copyright violations, and similar failures to provide source attribution. While it is likely that a generated text is drawn from multiple sources, genAI systems have produced verbatim text from a single

training source in a dataset (Quach, 2021). The problem may be more concerning with generation of images where elements selected from a dataset are combined. How can a user trust that genAI output does not violate copyright? Moreover, most genAI systems include some training texts taken without permission from copyrighted or otherwise protected sources. This may place users in a difficult and legally "gray" area when relying on materials generated by a genAI system.

## The Socio-Technical Context of GenAI

When considering the trustworthiness of a genAI system, as with any technology, it is important to focus not just on the model itself but also on those individuals, corporations, institutions, state actors, and others who are shaping it or incorporating it into their operations (Knowles & Richards, 2021; Lee & See, 2004). Technologies, including genAI, work as parts of larger socio-technical systems that bind together designers, users, material or digital artifacts, funding organizations, and myriad entities necessary for the emergence, deployment, functioning, maintenance, and definition of new technology. Thus, to understand a new technology's strengths, limitations, and biases—both generally and regarding specific use situations—and to begin to assess trustworthiness, the relevant socio-technical system must be considered. A focus on technical artifact misses examination of the requirements placed upon it by its stakeholders and other system components.

An example is the use of facial-recognition AI systems by police and other security agencies (Benjamin, 2019; Hill, 2020). Because genAI is trained on available datasets containing mostly white faces, it can distinguish among white faces more successfully than Black or Asian faces. As a result, its use has produced false identifications of innocent Black individuals (and others). While not a new phenomenon, the introduction of a presumably objective and error-resistant technology into a public institution known for racial discrimination meant that the misidentifications were rarely challenged and proved difficult to dispute. Many agencies continue to rely on facial-recognition technologies even though racial bias has been clearly demonstrated. The trustworthiness of genAI facial identification systems depends on holding police departments and agencies accountable for making well-reasoned decisions about how to use them. The genAI producers must also be required to design, build, and test their products to perform as advertised and avoid inequity in their application. Similarly, human resource genAI models are used to screen resumes despite racial and gender biases that cannot be corrected (Dastin, 2018). The interests of those deploying a genAI—often isolated to ease, cost, and availability—also impact its trustworthiness.

A major consideration when assessing the trustworthiness of a genAI system is financial interests behind the AI's development. Corporations are profit-driven, and selling products to as many customers as possible is their goal. Identifying the appropriate uses for their genAI model would reduce sales. In the absence of regulatory oversight, genAI companies must be accountable to the public. A genAI system can be employed by a company or government simply to avoid the cost of customer service, even if it introduces error, frustration, misinformation, and delay. An airline using genAI in customer service was held legally responsible for misinformation provided to its customers (Quach, 2024), and a government chatbot hosted by New York City advised business owners to break the law (Lecher, 2024). A company may use an AI to optimize profitability while disregarding worker health and safety.

Does the trust problem lie with the AI technology or with those who define the parameters of its use? Trustworthiness may best be considered a systemic

issue where the varied parts of the socio-technical network must be evaluated both individually and collectively (Knowles & Richards, 2021).

**Questions to consider about context of use for genAI include:**

- Who is profiting from the genAI and in what ways? How are their interests reflected in the way the genAI is used?
- What is the institutional context in which the genAI is being used? Who is determining what it will be allowed to decide or, checking on its determinations?
- Does the genAI reflect the same values as a human worker? Is it able to do that work better or is it just cheaper?
- How does the genAI change the dynamics for those interfacing with it?

## Hacking, Security, and GenAI

A final concern about the trustworthiness of genAI has received insufficient attention in system development, research, and media reporting: namely, security. How might genAI be used by malicious actors for gain? The possible hacking or "jailbreaking" through safety features of genAI impacts its trustworthiness in ways not yet known (Chakravorti, 2024; Afroogh et al., 2024). As genAI systems are embedded in more and more applications, their ability to cause harm increases if outside actors can use it for their own purposes. Vandalism of genAI has occurred through groups of users performing adversarial manipulation, such as encouraging Microsoft's Tay chatbot to spew racist comments (Victor, 2016.) Is it possible, for example, to prevent the hacking of a self-driving car or the news summaries a genAI produces? While genAI systems are trained on enormous databases and employ complex algorithms, systems have already been shown to be influenced by user behaviors. For example, in an early case of cyber vandalism, users altered a genAI chatbot's behavior through repeated interactions and caused it to respond with racist language (Lee, 2016). If genAI systems can't be trusted due to the possibility of hacking or jailbreaking, which can often be difficult to detect, can they be given responsibilities for information or decisions. What kinds of safeguards are possible, both for independent actors and for thwarting designed within genAI systems? Going forward, trust in genAI systems will require broad agreement on a set of ethical norms and regulatory systems (Floridi et al., 2018; Hagendorff, 2020; Spiegelhalter, 2020; Thiebes, Lins & Sunyaev, 2021.) For trustworthy genAI in context, ethical norms are needed within genAI systems to address these critical issues.

## Trustworthy GenAI Requires More Than "Humans in the Loop"

When the stakes are not high, trusting genAI's recommendations may be sensible even given the possibility that misinformation or partial information may be provided. The new technology of genAI systems draws quickly from a vast database of material, even though like most other information sources, it reflects mainstream Euro- American perspectives and corporate interests. However, if the stakes are consequential or vulnerability is significant, consulting other sources is recommended, as when considering medical diagnosis (Chakravorti, 2024). Additional sources may include other genAI systems, a person who is knowledgeable and trustworthy, or an expert. As a strategy for addressing uncertain information, using multiple sources to arrive at a consensus will increase the trustworthiness of the final result (Ross, 2024).

There are, however, situations where time, skills, cultural resources, personal ability, or institutional pressures limit the possibility of relying on multiple

sources. Continued technical developments improving the reliability and accuracy of genAI systems are critical. Reverse retrieval-augmented generation (reverse RAG) is one promising new approach where "the model extracts relevant information, then links every data point back to its original source content." (Plumb, 2025) GenAI models continually tested and refined to improve accuracy is another way to provide users with a measure of confidence in their output (Dwork and Minow, 2022).

Achieving trustworthy genAI will not be easy. To create a new ecosystem supporting trustworthiness requires changes in social, cultural, political, economic, and technological contexts:

- Incentives for producers to limit genAI applications to appropriate contexts of use.
- Defined legal liability for corporations or institutions making and/or employing genAI systems.
- Clear identification to the public when genAI models are being employed.
- Public support for enacting strict and enforceable genAI regulations and rights protections by governments.
- Community groups organized to share experiences and join activism around needs.
- Commercial channels for consumer feedback and redressing harms from genAI.

For situations where people are forced to use or are subjected to genAI, external regulation designed for the common good is required to protect individual users from arbitrary or misinformed determinations and the release or misuse of private data (Felzmann et al., 2019; Knowles & Richards, 2021; Schneier, 2023; Afroogh et al., 2024). Building accountability into the system is required to provide a recourse if it goes wrong, produces misinformation, or adversely affects them in order to increase genAI trustworthiness. Trust could be enhanced through transparency measures to promote user understanding of what the AI system is designed to accomplish and what sorts of factors it weights as most significant (Afroogh et al., 2024; Knowles & Richards, 2021; Zhou et al., 2023). One approach is to provide a supplier's declaration of conformity (SDoC) assuring that the AI system meets technical regulations and standards for safety and performance (Afroogh et al., 2024; Hind et al., 2019; Jacovi et al., 2021). Enforcement through legal or criminal liability for failures of an AI system, however, may be necessary to ensure wide compliance.

The consensus of scholars and activists is that trustworthy AI must embody several fundamental principles, including beneficence, non-maleficence, transparency and explicability, fairness, accuracy, auditability, accountability, environmental justice, social responsibility, and cultural sensitivity (Afroogh et al., 2024; Floridi et al., 2018; Hagendorff, 2020; Spiegelhalter, 2020; Thiebes, Lins & Sunyaev, 2021). But how will these values be represented throughout the design, development, application, maintenance, and deployment of AI systems?

Many recommendations suggest keeping "humans in the loop" (Charvorti, 2024), though human decisions are also imperfect (Aoki, 2021). Abbas (2019) suggests that AI systems might be designed with built-in triggers to automatically ask for human intervention. Human intervention to alter and even override genAI determinations may increase the trustworthiness of genAI in some situations and provide reassurance to users. However, a single human monitor alone may not be sufficient for averting serious harms as they occur. New roles in genAI for humans representing diverse stakeholders and backgrounds are required across every area of genAI model design, development, production, testing and use to ensure that human needs, values, and ethics are foregrounded. Some examples of roles for "humans in the loop" include:

- defining appropriate genAI applications for different sociocultural contexts
- gathering perspectives from representative public stakeholders affected by genAI
- designing relevant training data and task parameters to include ethical norms
- identifying representative data samples to include all those impacted by genAI
- testing systems before release to ensure equitable decisions across individuals and groups
- designing genAI interactions to better educate users about appropriate use
- responding to genAI user reports to identify harms as feedback for design

Focusing on human needs throughout the processes of design, production, testing, adoption, and use will improve the trustworthiness of genAI systems. It will also improve genAI success by increasing its understanding of meaning, flexibility in responding across contexts, equitable treatment of individuals, accuracy in performance, usefulness to users, and value to customers. With modest cost, providing rich human perspectives from diverse stakeholders "in the loop" will significantly improve the trustworthiness of genAI.

# VII. Conclusion

The possibility of building trustworthy and ethical genAI systems is real, but the challenges remain formidable. Generative AI presents profound opportunities and significant risks, shaping knowledge production and public trust in unprecedented ways. While genAI's capacity to synthesize vast data offers powerful tools for innovation, the opacity surrounding its data sources, decision processes, and accountability mechanisms significantly complicates assessments of trustworthiness. To mitigate these challenges, genAI systems must prioritize transparency, legal accountability, educational initiatives, and robust human oversight. Embracing "hopeful skepticism," users, policymakers, and developers alike can harness genAI's potential while safeguarding against its risks, ensuring these powerful technologies serve public interests and reinforce, rather than erode, societal trust.

# VIII. References

Abbass, H. A. (2019). Social integration of artificial intelligence: Functions, automation allocation logic and human-autonomy trust. *Cognitive Computation, 11*, 159–171. https://doi.org/10.1007/s12559-018-9619-0

Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024, March 12). *Trust in AI: Progress, challenges, and future directions*. arXiv. https://doi.org/10.48550/arXiv.2403.14680

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain. *Information Fusion, 99*, Article 101805. https://doi.org/10.1016/j.inffus.2023.101805

Altay, S., Berriche, M., & Acerbi, A. (2023). Misinformation on misinformation: Conceptual and methodological challenges. *Social Media + Society, 9*(1), 1–13. https://doi.org/10.1177/20563051221150412

Angwin, J., & Larson, J. (2016, May 23). *Machine bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Aoki, N. (2021). The importance of the assurance that "humans are still in the decision loop" for public trust in artificial intelligence: Evidence from an online experiment. *Computers in Human Behavior, 114*, Article 106572. https://doi.org/10.1016/j.chb.2020.106572

Arechar, A. A., Allen, J., Berinsky, A. J., Cole, R., Epstein, Z., Garimella, K., Gully, A., Lu, J. G., Ross, R. M., Stagnaro, M. N., Zhang, Y., Pennycook, G., & Rand, D. G. (2023). Understanding and combatting misinformation across 16 countries on six continents. *Nature Human Behaviour, 7*, 1502–1513. https://doi.org/10.1038/s41562-023-01641-2

Arnett, C. (2024). Dystopian dreams, utopian nightmares: AI and the permanence of racism. *The Georgetown Law Journal, 112*(6), 1299–1342.

Augenstein, I., Baldwin, T., Cha, M., Chakraborty, T., Ciampaglia, G. L., Corney, D., DiResta, R., Ferrara, E., Hale, S., Halevy, A., Hovy, E., Ji, H., Menczer, F., Miguez, R., Nakov, P., Scheufele, D., Sharma, S., & Zagni, G. (2024). Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence, 6*, 852–863. https://doi.org/10.1038/s42256-024-00881-z

Bahak, H., Taheri, F., Zojaji, Z., & Kazemi, A. (2023). *Evaluating ChatGPT as a question answering system: A comprehensive analysis and comparison with existing models.* arXiv. https://doi.org/10.48550/arXiv.2312.07592

Bakiner, O. (2023). Pluralistic sociotechnical imaginaries in artificial intelligence (AI) law: The case of the European Union's AI Act. *Law, Innovation and Technology, 15*(2), 558–582. https://doi.org/10.1080/17579961.2023.2245675

Banerjee, S., Agarwal, A., & Singla, S. (2024). *LLMs will always hallucinate, and we need to live with this.* arXiv. https://doi.org/10.48550/arXiv.2409.05746

Barocas, S. (2018, November 30). Accounting for artificial intelligence: Rules, reasons, rationales [Lecture]. Human Rights, Ethics, and Artificial Intelligence, Harvard Kennedy School Carr Center for Human Rights Policy.

Basu, K., & Alafriz, O. (2025, February 7). *AI's racial bias claims tested in court as US regulations lag. Bloomberg Law.* https://news.bloomberglaw.com/artificial-intelligence/ais-racial-bias-claims-tested-in-court-as-us-regulations-lag

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the New Jim Code*. Polity Press.

Block, N. (2024, December 29). Consciousness, artificial intelligence, and the philosophy of mind [Video]. YouTube. https://www.youtube.com/watch?v=wM1fcZr0iSk

Brandeis University Library. (n.d.). Citing generative AI: APA. https://guides.library.brandeis.edu/citeai/apa

Brewster, J., Wang, M., & Palmer, C. (2023, August). Plagiarism-bot? How low-quality websites are using AI to deceptively rewrite content from mainstream news outlets. *NewsGuard*. https://www.newsguardtech.com/misinformation-monitor/august-2023/

Burke, G., & Schellmann, H. (2024, October 26). *Researchers say an AI-powered transcription tool used in hospitals invents things no one ever said.* Associated Press. https://apnews.com/article/90020cdf5fa16c79ca2e5b6c4c9bbb14

Carroll, C. (2024, February 15). *AI-generated misinformation is everywhere. ID'ing it may be harder than you think. Maryland Today.* https://today.umd.edu/ai-generated-misinformation-is-everywhere-iding-it-may-be-harder-than-you-think

Carson, J. (2025, March 9). *ChatGPT response to query*. OpenAI.

Center for Countering Digital Hate. (2024, December 11). The double-edged sword of AI: How generative language models like Google Bard and ChatGPT pose a threat to countering hate and misinformation online. *Harvard Data Science Review*, (Special Issue 5). https://doi.org/10.1162/99608f92.be4e28f0

Chakravorti, B. (2024, May 3). AI's trust problem. *Harvard Business Review*. https://hbr.org/2024/05/ais-trust-problem

Chase, Z., Moran, S., & Yehudayoff, A. (2023). *Replicability and stability in learning*. Foundations of Computer Science (FOCS) Conference. https://doi.org/10.48550/arXiv.2304.03757

Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J. L., Clowez, G., Boileau, P., & Ruetsch-Chelli, C. (2024). Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research, 26*, Article e53164. https://doi.org/10.2196/53164

Chin, H., Lima, G., Shin, M., Zhunis, A., Cha, C., Choi, J., & Cha, M. (2023). User-chatbot conversations during the COVID-19 pandemic: Study based on topic modeling and sentiment analysis. *Journal of Medical Internet Research, 25*, Article e40922. https://doi.org/10.2196/40922

Crawford, K., & Paglen, T. (2019, September 19). Excavating AI: The politics of training sets for machine learning. https://excavating.ai

Dastin, J. (2018, October 10). *Amazon scraps secret AI recruiting tool that showed bias against women.* Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G/

Dunning, D., Fetchenhauer, D., & Schlösser, T. (2019). Why people trust: Solved puzzles and open mysteries. *Current Directions in Psychological Science, 28*(4), 366–371. https://doi.org/10.1177/0963721419838255

Dwork, C., & Minow, M. (2022). Distrust of artificial intelligence: Sources & responses from computer science & law. *Daedalus, 151*(2), 309–322. https://doi.org/10.1162/daed_a_01907

European Commission. (2019). *Ethics guidelines for trustworthy AI.* https://ec.europa.eu/futurium/en/ai-alliance-consultation

European Parliament. (2023). *EU AI Act: First regulation on artificial intelligence.* https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

European Union. (2024). *Key issues: EU AI Act*. https://www.euaiact.com/key-issue/5

Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society, 6*(1), 1–14. https://doi.org/10.1177/2053951719860542

Ferrucci, D., Levas, A., Bagchi, S., Gondek, D., & Mueller, E. T. (2013). Watson: Beyond Jeopardy!. *Artificial Intelligence, 199*, 93–105. https://doi.org/10.1016/j.artint.2012.06.009

Fletcher, R., & Nielsen, R. K. (2019). Generalised skepticism: How people navigate news on social media. *Information, Communication & Society, 22*(12), 1751–1769. https://doi.org/10.1080/1369118X.2019.1639065

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Future of Life Institute. (2017). *Asilomar AI principles.* https://futureoflife.org/ai-principles/

Gallup. (2023). *Historically low faith in US institutions continues.* https://news.gallup.com/poll/508169/historically-low-faith-institutions-continues.aspx

Germain, T. (2023, April 13). 'They're all so dirty and smelly:' Study unlocks ChatGPT's inner racist. *Gizmodo.* https://gizmodo.com/chatgpt-ai-openai-study-frees-chat-gpt-inner-racist-1850333646

Gigerenzer, G. (2023). Psychological AI: Designing algorithms informed by human psychology. *Perspectives on Psychological Science, 19*(5), 839–848. https://doi.org/10.1177/17456916231180597

Gillespie, N., Lockey, S., Curtis, C., Pool, J., & Akbari, A. (2023). Trust in artificial intelligence: A global study. The University of Queensland and KPMG Australia. https://doi.org/10.14264/00d3c94

GitHub. (n.d.). *Establishing trust in using GitHub Copilot.* https://resources.github.com/learn/pathways/copilot/essentials/establishing-trust-in-using-github-copilot/

Google Cloud. (n.d.). How Gemini for Google Cloud works. https://cloud.google.com/gemini/docs/discover/works

Google DeepMind. (2023). *GenCast: Learning skillful ensemble forecasting of medium-range weather.* https://deepmind.google/discover/blog/graphcast-ai-model-for-faster-and-more-accurate-global-weather-forecasting/

Google. (n.d.). Learn how Gemini in Gmail, Chat, Docs, Drive, Sheets, Slides, Meet & Vids protects your data. https://support.google.com/mail/answer/14615114

Gursoy, F., & Kakadiaris, I. A. (2022). Equal confusion fairness: Measuring group-based disparities in automated decision systems. 2022 IEEE International Conference on Data Mining Workshops (ICDMW), 137–146. https://doi.org/10.1109/ICDMW58026.2022.00025

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines, 30*, 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hancock, P. A., Kessler, T. T., Kaplan, A. D., Stowers, K., Brill, J. C., Billings, D. R., Schaefer, K. E., & Szalma, J. L. (2023). How and why humans trust: A meta-analysis and elaborated model. *Frontiers in Psychology, 14*, Article 1081086. https://doi.org/10.3389/fpsyg.2023.1081086

Haver, H. L., Ambinder, E. B., Bahl, M., Oluyemi, E. T., Jeudy, J., & Yi, P. H. (2023). Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. *Radiology, 307*(4), Article e230424. https://doi.org/10.1148/radiol.230424

Higgins, E. T. (1987). Self-discrepancy: A theory relating self and affect. *Psychological Review, 94*(3), 319–340. https://doi.org/10.1037/0033-295X.94.3.319

Hill, K. (2020, June 24). Wrongfully accused by an algorithm. *The New York Times Magazine*. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

Hind, M., et al. (2019). Increasing trust in AI through suppliers' declarations of conformity. *IBM Journal of Research and Development, 63*(4/5), 1–13. https://doi.org/10.1147/JRD.2019.2942286

Hon, L. C., & Grunig, J. E. (1999). Guidelines for measuring relationships in public relations. Institute for Public Relations. https://instituteforpr.org/measuring-relationships/

Horgan, L. (2022). The everyday of future-avoiding: Administering the data-driven smart city. *Information & Culture, 57*(2), 169–196. https://doi.org/10.7560/IC57203

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems, 43*(2), 1–55. https://doi.org/10.1145/3671011

Jacovi, A., Marasovic, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes, and goals of human trust in AI. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 624–635. https://doi.org/10.1145/3442188.3445923

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38. https://doi.org/10.1145/3571730

Jones, N. (2024). AI now beats humans at basic tasks—new benchmarks are needed, says major report. *Nature, 628*(8009), 700–701. https://doi.org/10.1038/d41586-024-01088-w

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., ... Kaplan, J. (2024). Language models (mostly) know what they know. arXiv. https://doi.org/10.48550/arXiv.2207.05221

Kamath, U., & Liu, J. (2021). *Explainable artificial intelligence: An introduction to interpretable machine learning*. Springer.

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). GPT-4 passes the bar exam. *Philosophical Transactions of the Royal Society A, 382*(2270). https://doi.org/10.1098/rsta.2023.0151

Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior, 160*, Article 108352. https://doi.org/10.1016/j.chb.2024.108352

Kumar, M., Mani, U. A., Tripathi, P., Saalim, M., & Roy, S. (2023, August 10). Artificial hallucinations by Google Bard: Think before you leap. *Cureus, 15*(8), Article e43313. https://doi.org/10.7759/cureus.43313

Kupferschmid, K. (2024, June 11). Transparency in copyright and artificial intelligence. Copyright Alliance. https://copyrightalliance.org/transparency-copyright-artificial-intelligence/

LaChapelle, & Tucker. (2023, November 28). Generative AI in political advertising. Brennan Center for Justice. https://www.brennancenter.org/our-work/research-reports/generative-ai-political-advertising

Lecher, C. (2024, March 29). NYC's AI chatbot tells businesses to break the law. The Markup. https://themarkup.org/news/2024/03/29/nycs-ai-chatbot-tells-businesses-to-break-the-law

Lee, D. (2016, March 25). Tay: Microsoft issues apology over racist chatbot fiasco. BBC News. https://www.bbc.com/news/technology-35902104

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Mahendrian, T., Moraes, G., Shamdas, M., Kern, C., Ledsam, J. R., Schmid, M. K., Balaskas, K., Topol, E. J., Bachmann, L. M., Keane, P. A., & Denniston, A. K. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *The Lancet Digital Health, 1*(6), e271–e297. https://doi.org/10.1016/S2589-7500(19)30123-2

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human-human and human-automation trust: An integrative review. *Theoretical Issues in Ergonomics Science, 8*(4), 277–301. https://doi.org/10.1080/14639220500337708

Maleki, N., Padmanabhan, B., & Dutta, K. (2024). AI hallucinations: A misnomer worth clarifying. *IEEE Conference on Artificial Intelligence (CAI)*. https://doi.org/10.1109/CAI59869.2024.00033

Marantz, A. (2024, March 11). Among the A.I. doomsayers. *The New York Times*. https://www.nytimes.com/2024/03/11/magazine/ai-doomsayers.html

McArthur, B. (2024, August 21). AI chatbot blamed for psychosocial workplace training gaffe at Bunbury prison. ABC News Australia. https://www.abc.net.au/news/2024-08-21/ai-chatbot-psychosocial-training-bunbury-regional-prison/104230980

Menczer, F., Crandall, D., Ahn, Y. Y., & Kapadia, A. (2023). Addressing the harms of AI-generated inauthentic content. *Nature Machine Intelligence, 5*, 678–680. https://doi.org/10.1038/s42256-023-00690-w

Merchant, B. (2025, May 12). So the LA Times replaced me with an AI that defends the KKK. Blood in the Machine. https://www.bloodinthemachine.com/p/so-the-la-times-replaced-me-with

Merken, S. (2025, February 18). AI 'hallucinations' in court papers spell trouble for lawyers. Reuters. https://www.reuters.com/technology/artificial-intelligence/ai-hallucinations-court-papers-spell-trouble-lawyers-2025-02-18/

Mihalcea, R., Ignat, O., Bai, L., Borah, A., Chiruzzo, L., Jin, Z., Kwizera, C., Nwatu, J., Poria, S., & Solorio, T. (2025). Why AI is WEIRD and shouldn't be this way: Towards AI for everyone, with everyone, by everyone. *Proceedings of the AAAI Conference on Artificial Intelligence, 39*(27), 28657–28670.

Milmo, D. (2023, February 9). *Google AI chatbot Bard sends shares plummeting after it gives wrong answer. The Guardian.* https://www.theguardian.com/technology/2023/feb/09/google-ai-chatbot-bard-error-sends-shares-plummeting-in-battle-with-microsoft

Molnar, C., Casalicchio, G., & Bischi, B. (2020). Interpretable machine learning: A brief history, state-of-the-art and challenges. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 417–431. Springer.

Nadella, S. (2016, June 28). Microsoft's CEO explores how humans and AI can solve society's challenges—together. *Slate.* https://slate.com/technology/2016/06/microsoft-ceo-satya-nadella-humans-and-a-i-can-work-together-to-solve-societys-challenges.html

National Academies of Sciences, Engineering, and Medicine. (2015). Trust and confidence at the interfaces of the life sciences and society: Does the public trust science? A workshop summary. *National Academies Press.* https://doi.org/10.17226/21809

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

National Telecommunications and Information *Administration*. (2023). Liability rules and standards. https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report/using-accountability-inputs/liability-rules-and-standards

Neeley, L. (2013, August 12). What the science tells us about 'trust in science'. COMPASSblogs. http://compassblogs.org/blog/2013/08/12/trust-in-science/

Nicoletti, L., & Bass, D. (2023, June 14). Humans are biased. Generative AI is even worse. Bloomberg Technology + Equality. https://www.bloomberg.com/graphics/2023-generative-ai-bias

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

OpenAI. (2023). *GPT-4*. https://openai.com/index/gpt-4-research/

OpenAI. (n.d.). How ChatGPT and our foundation models are developed. Retrieved February 18, 2025, from https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed

OpenAI. (n.d.). Sharing & publication policy. https://openai.com/policies/sharing-publication-policy/

Palmer, K., & Ross, C. (2024, November 19). Generative AI in focus as FDA's digital health committee meets. STAT News. https://www.statnews.com/2024/11/19/fda-digital-health-advisory-committee-artificial-intelligence/

Patel, A., & Sattler, J. (2023). Creatively malicious prompt engineering [Technical Report]. WithSecure Labs.

Pichai, S., & Hassabis, D. (2023, December 6). Introducing Gemini: Our largest and most capable AI model. Google Blog. https://blog.google/technology/ai/google-gemini-ai/

Plumb, T. (2025, May 11). Mayo Clinic's secret weapon against AI hallucinations: Reverse RAG in action. VentureBeat. https://venturebeat.com/ai/mayo-clinic-secret-weapon-against-ai-hallucinations-reverse-rag-in-action/

Quach, K. (2020, December 8). Uni revealed it killed off its PhD-applicant screening AI – just as its inventors gave a lecture about the tech. The Register. https://www.theregister.com/2020/12/08/texas_compsci_phd_ai/

Quach, K. (2021, March 18). What happens when your massive text-generating neural net starts spitting out people's phone numbers? The Register. https://www.theregister.com/2021/03/18/openai_gpt3_data/

Quach, K. (2024, February 15). Air Canada must pay damages after chatbot lies to grieving passenger about discount. The Register. https://www.theregister.com/2024/02/15/air_canada_chatbot_fine/

Quinn, B., & Milmo, D. (2024, November 26). How the far right is weaponising AI-generated content in Europe. The Guardian. https://www.theguardian.com/technology/2024/nov/26/far-right-weaponising-ai-generated-content-europe

Rahn, W. M., & Transue, J. E. (1998). Social trust and value change: The decline of social capital in American youth, 1976–1995. *Political Psychology, 19*(3), 545–565. https://doi.org/10.1111/0162-895X.00116

Resnick, D. B. (2011). Scientific research and the public trust. *Science and Engineering Ethics, 17*(3), 399–409. https://doi.org/10.1007/s11948-010-9210-x

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Roose, K. (2025, March 14). Powerful AI is coming. We're not ready. *The New York Times*.

Ross, A. (2024). AI and the expert: A blueprint for the ethical use of opaque AI. *AI & Society, 39*, 925–936. https://doi.org/10.1007/s00146-022-01564-2

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x

Saracini, C., Cornejo-Plaza, M. I., & Cippitani, R. (2025). Techno-emotional projection in human–genAI relationships: A psychological and ethical conceptual perspective. *Frontiers in Psychology, 16*, Article 1662206. https://doi.org/10.3389/fpsyg.2025.1662206

Salva. (n.d.). How GitHub Copilot handles data. https://resources.github.com/learn/pathways/copilot/essentials/how-github-copilot-handles-data/

Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors, 58*(3), 377–400. https://doi.org/10.1177/0018720816634228

Schneier, B. (2023, November 27). AI and trust. Belfer Center for Science and International Affairs, Harvard Kennedy School.

Spiegelhalter, D. (2020). Should we trust algorithms? *Harvard Data Science Review, 2*(1). https://doi.org/10.1162/99608f92.3b37f439

Strain, M. R. (2025). The case for AI optimism. *National Affairs, 63* (Spring).

Sutcliffe, A., Dunbar, R., Binder, J., & Arrow, H. (2012). Relationships and the social brain: Integrating psychological and evolutionary perspectives. *British Journal of Psychology, 103*(2), 149–168. https://doi.org/10.1111/j.2044-8295.2011.02061.x

Suzgun, M., Gur, T., Bianchi, F., Ho, D. E., Icard, T., Jurafsky, D., & Zou, J. (2024). Belief in the machine: Investigating epistemological blind spots of language models. arXiv. https://doi.org/10.48550/arXiv.2410.21195

Swenson, A., & Chan, K. (2024, March 14). Election disinformation takes a big leap with AI being used to deceive worldwide. Associated Press. https://apnews.com/article/artificial-intelligence-elections-disinformation-chatgpt-bc283e7426402f0b4baa7df280a4c3fd

Technology Innovation Institute. (2024). Home. https://www.tii.ae/

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets, 31*, 447–464. https://doi.org/10.1007/s12525-020-00441-x

Trust in American Institutions Challenge. (2024). Home. https://trust.leverforchange.org/submit

University of Michigan Medical School. (2025). AI tool label. https://medresearch.umich.edu/labs-departments/labs/TIERRA/our-work/ai-healthcare-label

Victor, D. (2016, March 24). Microsoft created a Twitter bot to learn from users. It quickly became a racist jerk. *The New York Times*. https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html

Wanless, A., Lai, S., & Hicks, J. (2025, February 20). Assessing national information ecosystems. Carnegie Endowment for International Peace. https://carnegieendowment.org/research/2025/02/assessing-national-information-ecosystems

Weichselbraun, A., Galvin, S. S., & McKay, R. (2023). Introduction: Technologies and infrastructures of trust. *The Cambridge Journal of Anthropology, 41*(2), 1–14. https://doi.org/10.3167/cja.2023.410202

Xiang, C. (2023, February 8). People are 'jailbreaking' ChatGPT to make it endorse racism, conspiracies. *Vice*. https://www.vice.com/en/article/people-are-jailbreaking-chatgpt-to-make-it-endorse-racism-conspiracies/

Yang, X., Song, B., Chen, L., Ho, S. S., & Sun, J. (2025). Technological optimism surpasses fear of missing out: A multigroup analysis of presumed media influence on generative AI technology adoption across varying levels of technological optimism. *Computers in Human Behavior, 162*, Article 108466. https://doi.org/10.1016/j.chb.2024.108466

Zaki, J. (2024). Hope for cynics: The surprising science of human goodness. Grand Central Publishing.

Zhou, J., Zhang, Y., Luo, Q., Parker, A., & Choudhury, M. (2023). Synthetic lies: Understanding AI-generated misinformation and evaluating algorithmic and human solutions. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3580971

Zhou, L., et al. (2024). Larger and more instructible language models become less reliable. *Nature, 634*, 61–68. https://doi.org/10.1038/s41586-024-07930-y

Zhu, L., Mou, W., Hong, C., Yang, T., Lai, Y., Qi, C., Lin, A., Zhang, J., & Luo, P. (2024). The evaluation of generative AI should include repetition to assess stability. *JMIR mHealth and uHealth, 12*, Article e57978. https://doi.org/10.2196/57978