

GENERATIVE AI, TRUST AND TRANSPARENCY IN AN AGE OF ALGORITHMIC KNOWLEDGE

Anjali Dave^{1*}

1. Assistant Professor, Faculty of Education, Waymade College of Education, The Charutar Vidya Mandal (CVM) University, Vallabh Vidyanagar, Gujarat, India

Abstract

Generative artificial intelligence (AI) is increasingly used in education, research, health, professional decision-making, public administration and communication, and it poses new challenges of trust, reliability, transparency and accountability. This article explores the interplay between trust and transparency in generative AI, focusing on the conditions that allow human users to rely on AI-generated outputs appropriately. The analysis distinguishes trust from trustworthiness and from reliance, and it treats generative AI as a sociotechnical and epistemic system rather than a purely technical one. It examines several dimensions of transparency – system-level, output-level, data, epistemic and interactional transparency – and highlights the shortcomings of disclosure, interpretability and explainability as stand-alone tools for building trust. The discussion suggests that when users cannot evaluate the sources, evidential support, limitations or origins of generated information, the ease of conversational interaction and the apparent authority of the output may lead to overconfidence and epistemic overreliance. The study concludes that trustworthy generative AI is not about maximising trust or disclosure but about enabling calibrated reliance. Technical reliability needs to be accompanied by epistemic and functional transparency, so that generative AI can be used in knowledge practices without assuming the status of an epistemic authority.

Keywords

Generative Artificial Intelligence, Trust, Transparency, Epistemic Trust, Algorithmic Knowledge, Provenance, Explainability, Accountability

1. Introduction

The spread of generative AI has changed human-computer interaction in ways that go beyond traditional concerns about automation. Whereas earlier AI systems were mainly used for classification, prediction and optimisation within fairly limited domains, generative AI creates language, images, software code, sound and other content that increasingly resembles human intellectual and creative production. Its growing use in education, research, healthcare, business, public administration, law and everyday communication raises a fundamental question: when should humans trust a plausible knowledge system that has neither human knowledge nor human responsibility? This question

cannot be answered simply by asking whether such systems are accurate. Reliance is not the same as performance, and it cannot be assured merely by making systems faster. A system may be fluent and convincing in some situations but not in others: it may answer simple questions correctly yet produce unsupported citations, explanations or information, or give incorrect information when used beyond its scope.

Students use large language models (LLMs) to explain concepts and create academic content; researchers use them to summarise literature and assist with writing; professionals seek support in drafting legal documents, analysing information and making recommendations; and users of conversational systems often treat these platforms as

*Corresponding author: anjali.dave123@gmail.com

knowledge banks. In all these scenarios, users must decide when to accept, question, verify or reject AI-generated output. These judgements depend not only on technical reliability but also on what users know about the system – what it can and cannot do, where its outputs come from and the conditions under which it operates. Trust is therefore a relational concept arising from the interplay between people, technology and the institutions in which technology is embedded. According to Lee and See (2004), the central issue in human–automation interaction is not simply establishing trust but forming appropriate reliance: users may over-trust and accept wrong results, or under-trust and reject a correct system. This distinction is significant for generative AI because its conversational interface may lead users to engage with it socially and cognitively rather than as a probabilistic computational system. The problem is therefore not confidence in generative AI as such, but whether the level and source of that confidence match the system’s capabilities and constraints. The challenge has intensified with the rise of LLMs, which produce coherent outputs across many areas without explicitly stating the uncertainty, origin or evidence behind them. Their apparent flexibility can create an illusion of knowledge, intentionality or reliability – a phenomenon that can be termed epistemic overreliance.

Bender et al. (2021) highlighted the dangers of ever-larger language models, including the risk that fluent language may be mistaken for understanding. Trust in generative AI therefore needs to be studied as a sociotechnical phenomenon rather than as a matter of individual psychological inclination or measurable system accuracy alone. When generative AI is integrated into widely used platforms, users may transfer their trust in the platform to the technology, even though they cannot assess the system’s architecture, training data, limitations or safety policies. This uneven distribution of knowledge between developers and users creates a dependency that makes transparency more crucial than ever.

Complete transparency, even if possible in theory, may be commercially constrained or undesirable where disclosure creates security risks. At the same

time, insufficient reporting on training data, model capabilities, evaluation methods or known risks can weaken accountability. The task is therefore not simply to call for full transparency but to establish what kind of transparency different actors need in order to make informed decisions and hold the relevant institutions accountable. This draws attention to the epistemological aspect of transparency: as generative AI systems become involved in creating and disseminating knowledge, the epistemic status of their products is often difficult to determine. Here, interpretability and explainability become relevant. Although the terms are often used interchangeably, interpretability typically refers to the understandability of a model’s internal functioning or behaviour, whereas explainability concerns the provision of reasons, representations or accounts that make a system’s outputs intelligible. Doshi-Velez and Kim (2017) called for a more consistent conceptualisation of interpretability and proposed methods for evaluating it. Transparency mechanisms that create only a façade of openness without allowing proper scrutiny can foster unwarranted confidence; trustworthy AI should therefore aim not at maximum disclosure in all situations but at meaningful disclosure suited to the risks and decision-making roles of different actors.

This reflects an emerging trend in AI governance, in which trustworthiness is seen as arising not only from the algorithm but also from organisational practices, human supervision, risk management, documentation and institutional responsibility. Regulatory developments also show a move towards differentiated transparency. The European Union’s Artificial Intelligence Act requires that people be informed when they interact with an AI system and that certain AI-generated or manipulated content be marked (European Union, 2024). Transparency is thereby linked to individuals’ ability to recognise that they are engaging with an AI system and to decide how far to depend on it, reflecting growing awareness that transparency must go beyond technical guarantees of model performance. Conversely, extensive organisational documentation may be of little use to those directly affected by AI-driven decisions or information.

Effective accountability requires a chain of responsibility running from the model developer and data provider through the deployer and institutional user to the person making decisions based on AI outputs. Generative AI complicates this chain because several actors and infrastructures may be involved in producing a single output: training data may come from various sources, a foundation model may be built by one organisation, fine-tuned by another and deployed in applications for entirely different purposes. Because generative models are trained on large and heterogeneous datasets, they also raise recurring questions of authorship, consent, attribution and the lawful use of creative and intellectual works. Users may not know which training data shaped the statistical patterns behind an output or how closely it resembles existing works, which makes responsible use difficult for users and problematic for copyright holders. In this sense, transparency concerns both information and the institutional conditions of content production.

Trust in generative AI should therefore not be viewed as a single outcome achieved through transparency. Rather, trustworthiness arises when several conditions are aligned: demonstrable reliability, understandable communication of limitations, appropriate human oversight, traceability and accountability, and institutional mechanisms for contestation and redress. Users should neither accept AI output blindly nor dismiss it hastily. Adopting a conceptual, literature-based approach, this article explores this sociotechnical and epistemological perspective on generative AI, trust and transparency. It proposes that the central question is not simply how to make generative AI systems more transparent but how to create forms of transparency that enable justified, appropriately calibrated reliance. It distinguishes technical transparency from epistemic transparency and institutional accountability, examines the limits of interpretability and explainability as tools of trust, and considers the role of disclosure, provenance, governance and regulation in fostering more responsible AI ecosystems.

2. Conceptualising Trust in Generative AI

Trust is one of the most frequently used and most ambiguous terms in discussions of AI. In public debate, trust in AI is often treated as a desirable goal, on the assumption that successful technological systems ought to generate more trust among users. This assumption is incomplete and problematic. The normative aim should not be to increase reliance on generative AI but to create conditions in which human reliance is proportionate to the system's actual abilities, limitations and risks. A highly trusted but unreliable system can be more dangerous than one that is not fully trusted but encourages cautious engagement (Yin et al., 2019). Studies of algorithmic advice also show that people may rely on algorithmic advice even without understanding how the algorithm works, suggesting that a sense of algorithmic authority can influence reliance independently of system performance (Logg et al., 2019).

2.1 Trust, Trustworthiness and Reliance

Trust, trustworthiness and reliance must be distinguished at the conceptual level, although they are often used interchangeably in discussions of AI. Trust is the attitude or expectation that an actor holds towards another actor or system in conditions of uncertainty or vulnerability. Trustworthiness refers to the attributes that make the trusted entity worthy of trust. Reliance is the behavioural outcome – the decision to depend on a system to carry out a task. These differences matter because a person may use a generative AI system without trusting it, or trust a system without relying on it. A researcher may believe that a generative AI system reliably produces summaries that are useful for most purposes, but not for reporting sensitive empirical results; conversely, a student may depend heavily on an AI tool for convenience while knowing that it sometimes gives wrong information.

Lee and See (2004), in a seminal work on trust in automation, proposed that the goal of automation design should be appropriate reliance rather than blind trust. Their framework is especially relevant to generative AI, where users often lack a clear understanding of the technologies that produced an outcome. In such situations, trust serves as a means of managing complexity and uncertainty, but when a system cannot be fully understood there is

potential for both overreliance and underreliance: the first occurs when users rely on automation beyond its capabilities, and the second when useful, proven automation is rejected or underused. The distinction between trust and trustworthiness therefore introduces an important normative principle: systems should not merely appear trustworthy but should be designed and managed to permit justified reliance, with attention to how the system actually behaves, its stability, how well it communicates uncertainty and the institutional structures in place when it fails.

Hoff and Bashir (2015) further show that trust in automation develops at several levels, distinguishing dispositional, situational and learned trust. Dispositional trust is a relatively stable individual tendency towards or away from trusting automation; situational trust is shaped by contextual factors such as the nature and consequences of the task; and learned trust develops through a person's own experience of a system's operation. This framework is useful for generative AI because users approach these systems differently, with judgements shaped by prior experience with technology, digital literacy, professional skills and the perceived importance of the task. Reliance also depends on observed failures: experimental studies show that people become overly negative towards algorithms after seeing them err, even when the algorithm has performed well in the past (Dietvorst et al., 2015). Conversely, a user who has had many satisfying interactions with a chatbot may become more trusting of it, and good performance in one area can lead to an inappropriate generalisation of competence. The system may then be trusted to interpret medical information, analyse legal requirements or generate scholarly citations on the basis of its success in writing routine messages. Such transfer of trust across domains is a major problem because the competence of current generative AI varies and depends on context (Kim et al., 2024).

2.2 From Interpersonal Trust to Human–Machine Trust

Conceptualising trust becomes more complex when the object of trust is a technological system rather than a human being. Classic theories of

interpersonal trust concern expectations of competence, integrity, benevolence and intention, and people can be held morally accountable for deception, negligence or betrayal. Generative AI systems, however, are not moral agents and do not have intentions in the traditional sense, so concepts of human trust cannot simply be transferred to machines. Caution is needed because conversational AI systems are often used in human-like ways, with user input resembling human interaction (Amershi et al., 2019): users pose questions, ask for advice, express uncertainty and receive responses in natural language.

LLM-powered chatbots may be perceived as a “quasi-other” and anthropomorphised by users despite being computational systems (Abercrombie et al., 2023; Heersmink et al., 2024). A chatbot appears friendly and open, as users can ask questions, seek clarification and receive instant answers, but this interactive accessibility does not mean that users understand how the system produced a response or whether the response is epistemically reliable. Conversational fluency can in fact mask the opacity of the underlying processes – an effect that may be described as anthropomorphic trust.

A further difficulty is that generative AI systems can simulate epistemic confidence without holding beliefs in the true sense. When a human expert makes a confident prediction, that confidence can be linked to expertise, evidence or experience, but when a generative system produces a confident statement, its language of certainty need not correspond to any justified degree of belief. The system can generate “I am certain” or “research clearly shows” without the epistemic commitment normally assumed in human communication. The trust relationship between humans and generative AI therefore differs from trust between humans. Users rely not only on an independent machine but also, usually implicitly, on the wider sociotechnical system around it, including the organisations that create the models, the institutions that deploy them, the data and computing resources on which they depend and the regulatory frameworks that govern their use.

2.3 Generative AI and the Problem of Epistemic

Trust

Epistemic trust – trust in information received from other people or external sources – is particularly useful for thinking about generative AI. Epistemic dependence is intrinsic to human knowledge: no individual can personally confirm all the scientific, historical, medical or legal information they encounter, and modern societies depend on expertise, institutions, publications, instruments and systems of verification. Generative AI is entering this epistemic ecosystem not only by processing information but also by presenting, synthesising and transforming it and generating knowledge claims. A user may ask a chatbot to describe a scientific theory, summarise a research debate, find literature or interpret a policy document; in such contexts the AI system acts as an intermediary between the user and the wider information environment, and the knowledge underlying AI-generated testimony is often hard to pinpoint.

Heersmink et al. (2024) argue that the opacity of both the data and the algorithm leaves users unable to know when LLMs are epistemically reliable. The natural, conversational flow of interaction further complicates the problem, as people may trust the content because it appears natural and fluent. Epistemic trust in generative AI should therefore not be considered binary – trust or distrust – but should acknowledge different contexts and kinds of epistemic dependence.

2.4 Trust as a Sociotechnical Relationship

A limitation of purely technical approaches to trustworthy AI is that they treat trust as a relationship between a person and an algorithm. In reality, generative AI is embedded in complex institutional settings. A user may not know who developed a specific model, how and on what datasets it was trained, what safety evaluations were performed or how the system is monitored after deployment, so trust is distributed among several actors. This relationship can be understood at three interdependent levels. System trust relates to the perceived reliability and competence of the technology. Organisational trust relates to the organisations that create and deploy the system – whether they accurately communicate system limitations, address known

risks, safeguard users' interests and act responsibly when failures occur. Institutional trust relates to the broader governance setting, including rules, standards, auditing procedures and complaint and appeal mechanisms. These levels can reinforce or undermine one another.

There is also a knowledge gap between developers and users. Technology companies may hold detailed information about model training, evaluation and deployment, whereas ordinary users interact only with the interface and outputs. This imbalance places users in a position of dependency in which they must rely on institutional statements about system capabilities. Modern societies routinely rely on complex systems that individuals cannot examine themselves, such as aircraft, pharmaceutical products, financial institutions and communication systems, and trust in these areas is reinforced through professional standards, regulation, certification, liability and independent evaluation. Individual user literacy, although useful, therefore cannot be the sole basis for responsible reliance on generative AI, since most people neither understand neural network architectures nor have time to evaluate training data before using an AI system.

3. Transparency as a Sociotechnical and Epistemic Problem

Generative AI involves several types of opacity. A system may reveal its structure without revealing its training data; it may cite sources without disclosing whether those sources actually informed the response; it may provide natural-language explanations without giving a faithful account of its computational processes; and an organisation may publish extensive documentation while withholding commercially sensitive information about model development and risk assessment. Transparency therefore cannot be understood as a single technical characteristic but as a multidimensional, sociotechnical phenomenon concerning the relationships among information, understanding, power, responsibility and the ability to assess or challenge technological outcomes. The main dimensions discussed in this section are summarised in Table 1.

Table 1. Dimensions of Transparency in Generative AI

<i>Dimension</i>	<i>Central question</i>	<i>Main limitation</i>
System-level transparency	What is the overall nature of the model (architecture, training methods, evaluation and restrictions)?	Technical detail may not help users judge a specific output.
Output-level transparency	What is the basis, source and uncertainty of a particular response?	Citations may be inaccurate or unrelated to the generated claim.
Data transparency	Where did the training data come from and how were protected materials used?	Outputs reflect learned statistical patterns, making source attribution difficult.
Epistemic transparency	Is the knowledge claim justified, and can it be verified by a third party?	Systems may present the appearance of evidence rather than its validity.
Interactional transparency	Does the conversational interface support genuine understanding?	Generated explanations may simulate, rather than reveal, the system's reasoning.

Source: Compiled by the author from the discussion in Section 3.

3.1 The Limits of Transparency as Disclosure

A widely held view in AI governance is that transparency can be increased simply by providing more information. This disclosure-based approach is intuitive: users should be told how a system functions, organisations should provide information about datasets whose sources are unknown, and explanations should be given when automated decisions affect a person. However, there is no automatic connection between information and transparency. Few people will read hundreds of pages of technical documentation, and a precise description of transformer architecture may be of little use to a teacher trying to decide whether an AI-generated explanation of a scientific phenomenon is correct.

Ananny and Crawford (2018) offer an important critique of the “transparency ideal” in algorithmic accountability, arguing that transparency should not be treated as a magic bullet for problems caused by complex algorithmic systems. Seeing into a system does not necessarily produce understanding of how it functions within its wider social and institutional context, because algorithms are situated in relationships among developers, organisations, datasets, infrastructures, users and regulatory institutions. Accountability therefore goes beyond the availability of technical information and requires awareness of the systems of responsibility surrounding them. This is especially important for generative AI. A user who receives a technical overview explaining

that an LLM predicts tokens by identifying patterns in a vast training set gains general knowledge about the technology, but this does not answer the epistemic question of why the system made a particular claim and what evidence supports it.

There is thus a difference between transparency of a system and transparency of its outputs. System-level transparency concerns the overall nature of the model, such as its architecture, training methods, evaluation techniques and restrictions, whereas output-level transparency concerns the basis or source of a specific response, its uncertainty and its reliability. Both are vital but serve different functions: a model card can provide useful information about a system's known limitations but cannot verify the facts in an individual output, while a citation in an AI-generated response may help verify a specific claim but says little about overall system performance or institutional risk management. The issue is therefore not only how much information is shared but whether the information disclosed helps other actors make better decisions.

3.2 Transparency for Whom and for What Purpose?

For disclosure to be meaningful, its audience and purpose must be identified. An AI researcher, policymaker, auditor, professional user and member of the public need very different information. A developer may need in-depth details of model architecture, training objectives and computational

performance; an independent auditor may require evaluation datasets, risk assessments, records of known failures and post-deployment monitoring; and a regulator may need information to assess compliance with legal requirements. An ordinary user most basically needs to know whether they are interacting with an AI system, what the system can reliably do, how personal data are used and when outputs should be verified. This points to a distinction between differentiated and universal transparency.

Not all information needs to be delivered in the same form to all actors, and indiscriminate disclosure can create its own problems, such as information overload, security risks and the exposure of legitimately protected intellectual property. The aim is not equal access to all information but access to the right information for meaningful oversight and informed reliance. Wachter et al. (2017) challenged simplistic notions of a universal right to explanation of automated decisions, showing that complex legal and conceptual issues cannot be resolved through the idea of full disclosure of algorithmic practices alone. Although their discussion concerned automated decision-making, it resonates strongly with generative AI.

3.3 Algorithmic Transparency and Technical Opacity

The most widely recognised form of opacity concerns the internal workings of computational models. Some machine-learning systems produce outputs through intricate mathematical relationships that are difficult to interpret, and this challenge is magnified by the scale of today's foundation models and their training processes, which cannot readily be explained in terms of simple cause-and-effect relationships. Burrell (2016) describes several forms of algorithmic opacity, and the explainable AI literature likewise treats the opacity of complex machine-learning systems as a key transparency concern, especially when users need to be meaningfully informed about how outputs are generated (Adadi & Berrada, 2018).

Opacity can be intentional, arising from business concerns: developers may disclose minimal information about training data, model weights,

optimisation methods or safety measures to protect intellectual property or for security reasons. Opacity can also result from technical illiteracy, since understanding current generative models requires specialist knowledge of machine learning, statistics, computational linguistics and large-scale infrastructure. Finally, intrinsic opacity refers to the difficulty of relating a specific output to an identifiable internal causal process in a complex model. No single transparency intervention can address all these forms. Publishing technical information may reduce intentional secrecy without significantly affecting intrinsic complexity or technical illiteracy, which has an important policy implication: making all the parameters of an LLM openly available would not allow an average user, regulator or even researcher to determine why a specific output was produced. Transparency must therefore go beyond the simplistic goal of "opening the black box"; the more pertinent goal is to establish mechanisms that ensure intelligibility, traceability and accountability despite technical complexity.

3.4 Data Transparency and the Problem of Provenance

For generative AI, transparency about data may be as important as transparency about algorithms. Large-scale models are trained on massive corpora drawn from books, the web, articles, software repositories, images, audio and other digital content, and the composition and sources of these datasets raise questions of copyright, representation, bias, privacy and informational integrity. Unlike traditional databases, most LLMs are not designed to retrieve and reproduce a particular document for each query; their outputs reflect learned statistical relationships distributed across model parameters, which makes simple source attribution difficult. However, this limitation should not be an excuse for a complete lack of transparency, and provenance will require more advanced solutions as the use of generative AI spreads. Users and institutions need information about the general origins and characteristics of the data environments from which models are developed, rights holders need ways of knowing how protected materials may have been used in training, and researchers need sufficient data to

evaluate limitations of representation and possible biases.

3.5 Epistemic Transparency and the Justification of Outputs

Epistemic transparency is one of the most important but least developed dimensions of generative AI governance. Whereas technical transparency asks how a system works, epistemic transparency asks how the knowledge claims it generates can be assessed. The difference matters because people engage with generative AI through its outputs, not its architecture. When a user requests an explanation of climate change, an interpretation of a historical event, a research summary or a recommendation for action, the central question is not how the tokens were predicted but whether the response is justified. Epistemic transparency therefore involves connecting a claim with its supporting evidence, its level of uncertainty and the possibility of third-party verification.

This requires AI systems and their infrastructures to give users meaningful signals of evidentiary grounding, such as source attribution, retrieval pathways, uncertainty indicators, knowledge cut-off information or explicit acknowledgement of evidentiary limitations. References generated within the system may be inaccurate, fabricated, out of date or unrelated to the claim being made, so transparency concerns not only the presence of sources but also the faithfulness of the relationship between generated claims and those sources. Apparently authentic citations, with plausible author names, journal titles and publication details, can give a false impression of academic legitimacy, and without traceability and accurate representation, citation itself becomes a form of performative authority. Responsible use of generative AI in knowledge production therefore requires helping users distinguish the appearance of citations from their validity.

3.6 Interactional Transparency and the Illusion of Understanding

The interface is another distinctive aspect of transparency in generative AI. Users can ask questions, seek clarification, challenge answers and engage in extended conversations, which can create a strong sense of transparency as the system responds

to their need for explanation. However, there is a fundamental difference between generating an explanation and accurately explaining the system's internal reasoning. An LLM can provide a comprehensive explanation of its response, but that explanation is itself another generated output, and there is no assurance of its accuracy unless it is validated independently against the computational processes that produced the original response. The explainability literature also shows that providing explanations does not make them meaningful and that their quality and effectiveness need to be systematically evaluated (Vilone & Longo, 2021). This creates a risk of explanatory simulation, in which the system generates explanations that meet human expectations of intelligibility without giving genuine access to the causal mechanisms underlying its behaviour. The distinction is important for trust, because a detailed explanation may be taken as evidence that the system's reasoning has been understood, whereas explanations do not necessarily produce appropriate reliance (Bansal et al., 2021).

The knowledge asymmetry between developers and users reinforces this problem. Developers decide which evaluation results are made public, how model limitations are explained, which information is treated as commercially sensitive and which risks are discussed publicly, while users typically encounter only a limited interface: they are expected to accept terms of service, engage with generated outputs and act on information they may be unable to verify independently. This poses a structural problem for trust, because users cannot simply choose to "not trust" a complex technological system that is integrated into educational, working and public structures.

3.7 The Paradox of Transparency

Efforts to increase transparency can produce paradoxical effects. An organisation may respond to calls for accountability with highly technical documentation that ordinary users cannot interpret; a model may offer increasingly detailed explanations that users mistake for genuine reasoning; excessive warnings may cause information overload and lead users to overlook critical conditions; and disclosure of a model's complexity may even produce a kind

of technological mystification in which users come to believe the system is too intricate to question. This paradox shows that the amount of information disclosed is not the best measure of transparency. In the AI4People ethical framework, Floridi et al. (2018) add explicability as a distinct principle alongside beneficence, non-maleficence, autonomy and justice, recognising that responsible AI must be both intelligible and accountable. Explainable AI research similarly places explainability within the wider context of trustworthy AI, together with robustness, transparency and accountability, rather than as a stand-alone technical feature (Ali et al., 2023). The concept is salient because it connects the capacity to understand AI systems with the ability to determine who is accountable for their outcomes. For generative AI, this implies a shift from informational transparency to functional transparency: informational transparency asks whether information has been disclosed, whereas functional transparency asks whether disclosure enables an actor to perform a meaningful function, such as assessing reliability, auditing risk, questioning a finding or holding a responsible party accountable.

4. Discussion

The analysis indicates that the key task with generative AI is not simply to gain users' trust but to create conditions for justified and appropriate reliance. Trust should therefore be considered a context-specific judgement rather than a general attitude towards technology, and the value of transparency depends not on the amount of information provided but on its ability to facilitate meaningful assessment, verification and accountability. Individual AI literacy alone cannot resolve the knowledge asymmetry between developers and users; effective institutional safeguards, clearly defined responsibilities and independent monitoring are also required. As generative AI spreads through education, research and professional decision-making, verifying sources, evaluating provenance and providing disclosure are likely to become integral to responsible knowledge production. Trustworthy generative AI requires that its outputs remain open to questioning, verification and appropriate contextualisation

rather than being accepted because of the convenience of the technology or the apparent authority of the source.

5. Conclusion

Generative AI has changed how people obtain and assess information. It generates fluent, contextually appropriate and apparently authoritative content, which opens new possibilities for knowledge work but also raises questions about the relationship between confidence and reliability. This analysis indicates that the appropriate response is neither blind faith nor blanket scepticism but appropriate reliance, which depends on the capabilities, limitations and consequences of each particular use. Transparency is central to this goal, but it should not be measured by how much information is provided; its value depends on whether users, professionals, auditors and institutions can use the information to assess outputs, identify limitations, adjudicate claims and establish responsibility when failures occur. Epistemic and functional transparency are particularly crucial for generative AI: information about how a system operates is important, but so is the capacity to determine the evidential basis and limitations of what the system produces.

There is no easy fix for trustworthy generative AI, whether through better model design or greater user awareness alone. It requires a convergence of technical reliability and meaningful disclosure with provenance, human oversight, organisational responsibility, institutional accountability and remedy processes. The goal should be to create conditions in which AI serves as an effective part of knowledge practices without being mistaken for an autonomous source of epistemic authority. In short, trust must be calibrated, transparency must be meaningful, and computational fluency must not stand in place of evidence.

References

- [1]. Abercrombie, G., Cercas Curry, A., Dinkar, T., Rieser, V., & Talat, Z. (2023). Mirages: On anthropomorphism in dialogue systems. In Proceedings of the 2023 Conference on Empirical

- Methods in Natural Language Processing (pp. 4776–4790). Association for Computational Linguistics.
- [2]. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160.
- [3]. 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 trustworthy artificial intelligence. *Information Fusion*, 99, Article 101805.
- [4]. Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). Association for Computing Machinery.
- [5]. Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973–989.
<https://doi.org/10.1177/1461444816676645>
- [6]. Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. S. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). Association for Computing Machinery.
- [7]. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery.
- [8]. Burrell, J. (2016). How the machine “thinks”: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 1–12.
- [9]. Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114–126.
- [10]. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning (arXiv:1702.08608). *arXiv*.
- [11]. European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.
- [12]. 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, 689–707.
- [13]. Heersmink, R., De Rooij, B., Clavel Vázquez, M. J., & Colombo, M. (2024). A phenomenology and epistemology of large language models: Transparency, trust, and trustworthiness. *Ethics and Information Technology*, 26(3), Article 41.
- [14]. Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3), 407–434.
- [15]. Kim, S. S. Y., Liao, Q. V., Vorvoreanu, M., Ballard, S., & Vaughan, J. W. (2024). “I’m not sure, but...”: Examining the impact of large language models’ uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 822–835). Association for Computing Machinery.
- [16]. Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80.

- [17]. Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151, 90–103.
- [18]. Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76, 89–106.
- [19]. Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2), 76–99.
- [20]. Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–12). Association for Computing Machinery.