Humanities & Social Sciences Communications

COMMENT

https://doi.org/10.1057/s41599-026-07513-4 OPEN

Check for updates

Plausibility, persuasion, and truth: why language models may appear designed to deceive

Giulio Vidotto[1]✉

Large language models can produce fluent, coherent, and persuasive responses even when the information on which they rely is partial, contested, or false. In disputed domains, this may leave some users with the impression that they are being deliberately misled. This Comment argues that the phenomenon is better understood in structural than intentional terms. It results from the convergence of four features of current systems: optimization for plausibility rather than truth, post-training incentives that reward helpful and persuasive answers, structural hallucination, and source bias rooted in asymmetries of knowledge production and digitization. These structural tendencies are further reinforced at the point of uptake by cognitive vulnerabilities such as automation bias and fluency-based truth effects. Recent evidence on conversational persuasion suggests that gains in persuasive force may come at the expense of factual accuracy. The governance problem, then, is not primarily to infer intent, but to identify the mechanisms through which epistemic distortion is produced. This Comment therefore proposes a minimal framework for epistemic auditing that distinguishes factual error, systematic omission, corpus bias, and post-training- or prompt-induced distortion, with a view to more discriminating oversight and clearer lines of responsibility.

[1] Department of General Psychology, University of Padua, Padova, Italy. ✉email: giulio.vidotto@unipd.it

Introduction

In a recent analysis of ChatGPT's usefulness in psychotherapy, Raile (2024) records a finding whose force lies, precisely, in the fact that it might at first seem almost banal: when asked about treatment options for social anxiety disorder, the model regularly recommended cognitive behavioral therapy, mindfulness-based approaches, and psychodynamic therapy, while passing over the wider range of established alternatives. Only when logotherapy was expressly mentioned in the prompt did it refer to Frankl's technique of paradoxical intention, and even then in partial form. Raile reads this pattern as an effect of training bias: the model reproduces the distribution of therapeutic approaches within the corpora on which it has been trained, where CBT occupies a position of predominance not by virtue of any deliberate editorial design, but because it is disproportionately represented in the digitized literature from which the training data are drawn. Users who possess no independent clinical expertise and who turn to the model for guidance thus receive an account of the field that is assured, fluent, and plausible, yet one that systematically misstates the actual diversity of therapeutic options. The practical harm is tangible; the mechanism, however, is structural.

This example is not some marginal curiosity. It discloses, rather, a more general pattern operating across domains and at differing scales: language models generate outputs shaped by the distribution of available text, while users lacking independent expertise may be inclined to receive that shaping as though it were neutral, or even authoritative. Where the relevant terrain is contested, whether historical events, geopolitical conflicts, therapeutic choices, or electoral information, the effect may in some circumstances be construed as deliberate manipulation. The recurrent omission of certain perspectives, the confident ratification of culturally dominant narratives, and the fluent reiteration of widely held falsehoods can together produce an experiential impression of agency or intention, even where no such intention has been established. The question addressed by this Comment is whether that impression is epistemically warranted, and what follows for governance if it is not.

The growing literature on trust in AI approaches cognate concerns from several directions, encompassing technical metrics of trustworthiness, user psychology, and frameworks of governance (Afroogh et al. 2024). Research on human-AI relations has likewise shown how the social affordances of language models, namely their fluency, their apparent responsiveness, and their capacity for personalization, shape user perceptions and conduct in ways that extend well beyond mere task performance (Andersson 2025; Kirk et al. 2025). This Comment does not aim to offer either a general theory of trust in AI or a comprehensive account of human-AI relational dynamics. Its concern is narrower. It seeks to explain why certain structural properties of language models may, under identifiable conditions, give rise to the experiential impression of intentional deception. What remains insufficiently specified in the existing literature is precisely this mechanistic connection between training dynamics, patterned epistemic distortion, and the appearance of deliberate intent. That, and not trust or relationality in their most general sense, is the gap with which this paper is concerned; and it is from that gap that its argument about governance proceeds.

This appearance of purposiveness is likely reinforced by two familiar features of human interpretation. First, human observers are often inclined to attribute agency or intention to coherent, goal-like, and recurrent patterns even where no intentional agent is demonstrably present (Heider and Simmel 1944; Marañes et al. 2024). Second, the conversational form of interaction with language models encourages users to interpret outputs against background expectations of cooperative exchange (Grice 1989). In this sense, fluency and apparent responsiveness do not merely convey information; they also supply cues that make structurally generated distortions easier to experience as purposive or misleading.

The argument can now be specified in four connected steps. First, this Comment describes the interacting mechanisms that generate outputs which may appear deceptive without presupposing deceptive intent: optimization for plausibility, post-training incentives oriented toward persuasion, structural hallucination, and source bias. Second, it shows that these mechanisms operate synergistically rather than merely cumulatively, producing a pattern at once coherent and directional, one that mimics intentional influence precisely because it recurs across topics and across users. Third, it considers how recent experimental evidence on conversational persuasion sharpens the problem, above all through the documented tension between persuasive capacity and factual accuracy. Fourth, building upon this account, it argues that the relevant question for governance is not how intention might be inferred, but how the mechanisms through which epistemic distortion is produced may be identified and audited. On that basis, it proposes a framework of epistemic auditing that distinguishes structural artifacts from intentional manipulation and renders responsibility traceable along the chain through which these systems are produced and deployed.

The intent-free reading advanced here does not diminish the harms that such effects may cause. For users who form distorted beliefs on the basis of asymmetrical or inaccurate model outputs, the experiential consequence remains the same whether the asymmetry were deliberately engineered or emerged from the structure of a training corpus. For regulators, providers, educators, and policymakers, however, the distinction between structural artifact and intentional manipulation is hardly a merely academic matter: it determines which interventions may plausibly prove effective, which expectations of mitigation remain realistic, and at what point in the chain responsibility may be meaningfully assigned.

The Mechanics of Apparent Deception

A standard language model is trained to predict the most probable continuation of a text given its context. Its objective function is not truth-tracking but distributional plausibility: the generation of sequences that resemble human discourse as it appears in available corpora. Post-training procedures, most prominently reinforcement learning from human feedback (Ouyang et al. 2022), can substantially modify system behavior (increasing instruction-following, reducing certain failure modes, shaping tone and register), but they do not respecify the system as a fact-checker or a truth-tracking device. A model trained to generate responses that human raters prefer is a model optimized for perceived quality. Perceived quality and epistemic quality overlap but are not identical, and the divergence between them is precisely where the appearance of deception is generated.

This asymmetry, by itself, would still not be enough. It becomes consequential only in interaction with three further mechanisms: structural hallucination, source bias, and the fluency-credibility relationship. The critical point is that these four factors do not add their effects; they amplify each other.

Hallucination can be defined as the production of content not supported by available data or context, presented with the surface features of accurate output. It is a structural risk of any system that must generate coherent, well-formed responses under the constraint of apparent completeness. Ji and colleagues (2023) provide a comprehensive taxonomy of hallucination types and the conditions that increase their probability across generation tasks. The psychologically central observation is that hallucinated

responses do not present themselves as uncertain or anomalous: they exhibit the same fluency, apparent confidence, and narrative coherence as accurate responses. Linguistic well-formedness becomes the vehicle of error. The benchmark results of Lin and colleagues (2022), using TruthfulQA, make the epistemic structure of this problem precise: models systematically generate plausible-sounding but false answers to questions where cultural stereotypes or widespread false beliefs provide a more "writable" response than the accurate one. The system does not lie; it imitates the texture of reliable discourse while being unconstrained by its content requirements.

Post-training optimization adds a further channel. Reinforcement learning from human preferences selects for outputs that human raters endorse, and human preference for AI outputs is substantially influenced by fluency, apparent confidence, and the appearance of helpfulness. Reber and Schwarz (1999) demonstrate that processing fluency directly increases truth judgments: what is easier to process is more likely to be judged as true. A language model optimized for human preference is in part a fluency maximizer; the perceived competence socially attributed to intelligent systems adds a credibility multiplier that operates independently of content accuracy. Users who receive confident, well-structured responses on topics they cannot independently verify are in a structurally vulnerable epistemic position, not because the system is adversarial, but because the same properties that make responses useful also make errors persuasive. Kadavath and colleagues (2022) show that some models, under specific elicitation conditions, can generate reasonably calibrated uncertainty signals: the capacity to communicate epistemic limitations exists. The problem is that this capacity is not reliably activated in ordinary conversational outputs and is typically not prompted by users who lack the technical vocabulary to request it.

Source bias closes the system from the input side. Training is anchored in available text, and availability is not a neutral property: it tracks editorial power, dominant languages, digitization infrastructure, archival priorities, and institutional capacity. What appears canonical or standard in a model's outputs on contested topics typically reflects what is most densely represented in global corpora, which in turn reflects asymmetries in the production and preservation of knowledge. Bender and colleagues (2021) identify the epistemic and social risks of training on large, under-documented corpora, including the difficulty of tracing provenance and assessing representativeness. More recent empirical work documents systematic patterns of ideological and geopolitical asymmetry across models developed in different national and institutional contexts (Buyl et al. 2026; Noels et al. 2026). These asymmetries are not the product of explicit choices by individual actors at the moment of inference. They are emergent properties of a supply chain that introduces selection pressures at each stage: who writes, who publishes, who archives, who digitizes, who indexes, and what enters the training data.

The interaction among these four mechanisms is what produces the experiential appearance of intent. No single mechanism is sufficient: plausibility optimization without post-training incentives produces less persuasive outputs; hallucination without fluency is more easily detected; and fluency without a broader pattern of output directionality produces noise rather than a consistent impression. The role of source bias becomes especially consequential once these tendencies are encountered by users under conditions that favour epistemic over-reliance, a point to which the discussion returns below. Together, they generate outputs that are simultaneously coherent, consistent across topics, and systematically skewed in ways that reflect training distributions and optimization targets. This combination, characterized by coherence, consistency, and directionality, is well-suited to eliciting agency attribution in human observers. Human interpreters are often inclined to read coherent and recurrent directional patterning as purposive, even where no such intention has been demonstrated. In that sense, the impression of intentional deception can be understood as a plausible cognitive response to a structural technical phenomenon.

Amplification: Persuasion, Cognition, and The Source Chain

The structural mechanisms described above do not operate in an epistemic vacuum. They interact with documented features of human cognition and with experimental evidence on models' persuasive capacity in ways that substantially amplify their effects on belief formation.

Automation bias, the tendency to over-weight automated system outputs relative to other sources of information (including in the presence of contradicting evidence), is well documented across applied settings (Parasuraman and Riley 1997). The systematic review by Goddard and colleagues (2012) identifies the conditions under which the bias is strongest: when the task domain is complex, when the user lacks independent expertise, and when the system presents its outputs with apparent confidence. When language models are consulted under such conditions, the epistemic risks associated with their outputs are likely to be amplified. The practical consequence is that, in high-complexity and low-expertise contexts, model errors may receive disproportionate epistemic weight. What the user receives is not just a response, but one that is more likely to be accepted precisely when the resources for critical evaluation are weakest.

The persuasive capacity of current models is not merely a theoretical inference from training dynamics. Bai and colleagues (2025) demonstrate that model-generated messages influence opinions on contested policy issues with effectiveness comparable to texts produced by humans. Salvi and colleagues (2025) show that GPT-4, when personalized, significantly exceeds human interlocutors in persuasive effectiveness during debate conversations. The most methodologically consequential result is that of Hackenburg and colleagues (2025): testing persuasion levers systematically across hundreds of political issues and multiple models, they find that techniques that increase persuasion (specific post-training and prompting strategies) can reduce factual accuracy. This relationship is not incidental. It reflects a structural trade-off: optimizing outputs for persuasive effect in domains where compelling claims are not always the most accurate ones produces a measurable cost to epistemic quality. If the operational objective is to convince, truth functions as a cost variable to be managed rather than a constraint to be satisfied.

The implications of this trade-off extend beyond the design choices of individual providers. They bear directly on deployment decisions. A system whose persuasive capacity is well documented, and whose deployment context involves opinion formation on contested issues (electoral choices, health decisions, historical interpretation, geopolitical assessment), is a system where the persuasion-accuracy trade-off is not a theoretical risk but a predictable operational feature. Deploying such a system in those contexts without appropriate safeguards is not a passive decision; it is a choice whose consequences are foreseeable.

Source bias functions as a third amplifier, operating at a different scale and through a different mechanism than cognitive bias but producing structurally similar effects on belief formation. The claim that source bias is pervasive does not require attributing deliberate falsification to any individual actor. It requires only acknowledging that the global infrastructure of knowledge production is asymmetric, and that asymmetry accumulates across successive stages of the supply chain. Noels and colleagues (2026) document how moderation and censorship practices vary systematically across models trained in different geopolitical

regions, with patterns reflecting the national and cultural priorities of providers. Buyl and colleagues (2026) identify measurable ideological disparities in how models from different regions evaluate political figures and events. These patterns are not anomalies in otherwise neutral systems; they are predictable consequences of training on corpora that are themselves products of asymmetric institutional and political economies of knowledge.

Having introduced source bias as an upstream feature of the training pipeline, the discussion now turns to its downstream amplification at the level of user uptake and deployment context.

Source bias and automation bias interact in governance-relevant ways. Users who encounter consistent, confident, fluently presented accounts of contested events may be cognitively predisposed to accept them, and the consistency of those accounts may function as apparent corroboration. In some cases, such consistency may reflect genuine epistemic convergence; in others, corpus asymmetry; in still others, some combination of the two. From the user's perspective, however, consistency alone may be insufficient to distinguish among these possibilities. A user who repeatedly receives Anglo-centric accounts of colonial conflicts, or model-generated summaries of geopolitical disputes that systematically reflect one party's framing, cannot from experience alone distinguish corpus asymmetry from editorial manipulation. The experiential result is identical in both cases: a confident, coherent, directional account that excludes or marginalizes certain perspectives while presenting itself as informative and balanced.

Epistemic Auditing: from Structural Diagnosis to Traceable Responsibility

The analysis in the preceding sections has a direct methodological implication. If the effects that users may attribute to intentional deception are products of interacting structural mechanisms, then the appropriate governance response should not depend primarily on adjudicating intent. It should begin, rather, with empirical diagnosis: identifying which mechanisms are operating, at what intensity, and in which deployment contexts. This does not imply that existing legal or regulatory frameworks uniformly hinge on proof of deceptive intent; the narrower point is that, in public and policy discourse, the language of deception often invites intentional readings that are analytically less useful than structural diagnosis. Without a diagnostic framework, regulatory and institutional responses risk addressing visible symptoms while leaving generative mechanisms intact. They also risk conflating structurally distinct failure modes that require different interventions or, conversely, using technical complexity as grounds for regulatory inertia.

By epistemic auditing I mean, at this stage, an initial set of repeatable and publicly verifiable measures meant to distinguish four operationally distinct classes of output failure: (a) factual error; (b) systematic omission; (c) corpus bias; and (d) bias induced by post-training or prompt design. This framework is intended as a diagnostic starting point rather than as a validated standard for cross-provider evaluation. These classes are analytically distinct and require different remedies. Factual error, i.e., the production of claims that are verifiably false, can be addressed through retrieval augmentation, grounding techniques, and calibration training. Systematic omission, i.e., the consistent failure to represent certain perspectives, methods, evidence, or populations, requires corpus diversity requirements and source transparency standards, as Raile's (2024) analysis of ChatGPT's therapeutic recommendations illustrates concretely: the omission of non-CBT approaches is not a factual error but a structural gap in representation. Corpus bias, i.e., the systematic skewing of outputs toward the perspectives most heavily represented in training data, requires interventions upstream of training: documentation standards, representation requirements, and institutional investment in multilingual and non-Western knowledge infrastructure. Bias induced by post-training or prompt design, i.e., the systematic modification of output content through reward model choices or prompting strategies, requires design transparency and usage policy standards. Conflating these classes produces interventions that are misspecified relative to their targets. The risk taxonomy of Weidinger and colleagues (2022), which distinguishes misinformation, manipulation, exclusion, and downstream harms, offers a principled basis for classification that the present framework extends toward operational measurement.

A minimal set of auditing criteria that could be compared across providers and deployment contexts includes five classes of measures: (a) the rate of verifiable claims, defined as the proportion of empirical assertions in model outputs that are accompanied by citable external evidence, which provides a baseline measure of output groundedness and distinguishes systems that generate text from systems that retrieve and synthesize documented knowledge; (b) the assertiveness-to-uncertainty ratio, defined as the frequency of confidence markers relative to hedging formulations, conditional statements, or explicit acknowledgments of epistemic limitation, which measures the accuracy with which models communicate the epistemic status of their outputs independently of propositional content (a system that generates accurate responses but fails to communicate uncertainty where uncertainty exists is structurally deceptive in a restricted sense that does not require false claims); (c) the institutional and linguistic diversity of sources, defined as the geographical, linguistic, and institutional distribution of materials explicitly or implicitly referenced in outputs on contested topics, which provides a proxy for corpus asymmetry in the domain of use; (d) sensitivity to prompt variations, defined as the degree to which output content changes when the same question is posed in different registers, languages, or framings, which identifies outputs whose apparent objectivity is an artifact of prompt conditions rather than a property of the information itself; and (e) coherence under follow-up questioning, defined as the degree to which outputs remain consistent when users pose verification questions, request evidence, or challenge specific claims, which distinguishes systems that track an underlying representation from systems that generate locally coherent text without global consistency.

These measures are not benchmarks for reassurance. A model that shows a high rate of unverifiable claims, low source diversity, high sensitivity to prompt framing, and instability under follow-up questioning should not be deployed for opinion formation in electoral, judicial, or educational contexts without external controls, regardless of its performance on standard benchmark tasks. Making these metrics public and, where possible, standardized would begin to transform auditing from an internal quality check into a tool for differentiated governance: a technical basis for disclosure obligations, usage restrictions proportionate to epistemic consequences, and deployment standards linked to measurable properties of system behavior. The present Comment does not empirically validate these measures across models or deployment settings; it reconstructs, on the basis of available evidence, a framework for distinguishing mechanisms that should be tested comparatively in future work.

The auditing framework also clarifies the basis of provider responsibility without making proof of manipulative intent the hinge variable. The persuasion-accuracy trade-off documented

by Hackenburg and colleagues (2025) is now part of the public scientific record. When a provider, with access to this evidence, continues to optimize systems for persuasive effectiveness in contexts where accuracy is critical, and deploys them in those contexts without appropriate disclosures or safeguards, the resulting distortions are no longer plausibly treated as unforeseeable artifacts. Awareness of structural risk has a constitutive effect on the normative status of subsequent choices: it shifts the relevant question from intent to risk management under actual control.

This logic extends along the deployment chain. Platforms that integrate language models into high-stakes information flows without contextualization share responsibility for predictable consequences of that integration. Regulators who fail to mandate auditing standards despite available evidence share responsibility for the governance gap that remains. Educational institutions that deploy these systems without critical literacy training share responsibility for the epistemic vulnerability they leave unaddressed.

The chain of responsibility implied by this framework is longer than a simple provider-user model suggests, but it is also more operational. The alternative, attributing diffuse responsibility to an opaque technological system in which no actor is identifiable in virtue of competence and control, tends to make responsibility practically unassignable. A distributed allocation, proportional to demonstrated competencies and actual control over the mechanisms that generate risk, is both more equitable and more practically enforceable.

Conclusion

If a language model gives a user the impression of deliberate deception, the fitting response is to treat that impression as a psychological datum and a technical hypothesis, not as a moral verdict. The four mechanisms analyzed in this paper (plausibility optimization, post-training persuasion incentives, structural hallucination, and source bias) are sufficient, taken together, to account for the observed effects without invoking agency. Their interaction produces the coherence, consistency, and directionality that can elicit agency attribution in human observers; the inference of intent thus appears as an intelligible response to a structural phenomenon, even where no such intent has been shown.

The distinction between structural artifact and intentional manipulation is consequential because it determines the intervention space. Sanctioning a provider for corpus bias, by itself, does not reduce corpus bias; source-transparency requirements and independent auditing are more likely to do so. Expecting neutrality on contested historical or geopolitical topics is technically naive; requiring measurable source diversity, calibrated uncertainty communication, and deployment standards proportionate to epistemic risk is both technically feasible and politically legitimate. The average user who does not know what reinforcement learning from human feedback is will continue to be exposed to the persuasive effects of fluent, confident, plausible text. But an informed governance ecosystem can build mitigation architectures that make risk measurable, responsibility traceable, and harm governable. In the absence of any absolute neutrality, which is neither achievable nor coherently definable, this is probably the furthest a realistic ambition can go.

Data availability

Data sharing is not applicable to this research as no data were generated or analyzed.

Received: 4 March 2026; Accepted: 27 April 2026;
Published online: 09 May 2026

References

Afroogh S, Akbari A, Malone E et al. (2024) Trust and distrust in AI: Progress, challenges, and future directions. Humanities Soc Sci Commun 11:1512. https://doi.org/10.1057/s41599-024-04044-8

Andersson G (2025) AI companionship: Emotional fast food?. Humanities Soc Sci Commun 12:726. https://doi.org/10.1057/s41599-025-05536-x

Bai H, Voelkel JG, Muldowney S et al. (2025) LLM-generated messages can persuade humans on policy issues. Nat Commun 16: 5582. https://doi.org/10.1038/s41467-025-61345-5

Bender EM, Gebru T, McMillan-Major A et al. (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 (FAccT '21). Association for Computing Machinery, New York, NY, USA, pp. 610–623. https://doi.org/10.1145/3442188.3445922

Buyl M, Rogiers A, Noels S et al. (2026) Large language models reflect the ideology of their creators. Npj Artif Intell 2:7. https://doi.org/10.1038/s44387-025-00048-0

Goddard K, Roudsari A, Wyatt JC (2012) Automation bias: A systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 19(1):121–127. https://doi.org/10.1136/amiajnl-2011-000089

Grice HP (1989) Logic and conversation. In: Studies in the way of words. Harvard University Press, Cambridge, MA, pp. 22–40

Hackenburg K, Tappin BM, Hewitt L et al. (2025) The levers of political persuasion with conversational artificial intelligence. Science. 390(6777). https://doi.org/10.1126/science.aea3884

Heider F, Simmel M (1944) An experimental study of apparent behavior. Am J Psychol 57(2):243–259. https://doi.org/10.2307/1416950

Ji Z, Lee N, Frieske R et al. (2023) Survey of hallucination in natural language generation. ACM Comput Surv 55(12):1–38. https://doi.org/10.1145/3571730

Kadavath S, Conerly T, Askell A et al. (2022) Language models (mostly) know what they know. arXiv:2207.05221. https://doi.org/10.48550/arXiv.2207.05221

Kirk HR, Wachter S, Mittelstadt B et al. (2025) Why human-AI relationships need socioaffective alignment. Humanities Soc Sci Commun 12:582. https://doi.org/10.1057/s41599-025-04532-5

Lin S, Hilton J, Evans O (2022) TruthfulQA: Measuring how models mimic human falsehoods. In: Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers). Association for Computational Linguistics, Dublin, Ireland, pp. 3214–3252. https://doi.org/10.18653/v1/2022.acl-long.229

Marañes C, Gutierrez D, Serrano A (2024) Revisiting the heider and simmel experiment for social meaning attribution in virtual reality. Sci Rep 14(1):17103. https://doi.org/10.1038/s41598-024-65532-0

Noels S, Bied G, Buyl M et al. (2026) What large language models do not talk about: An empirical study of moderation and censorship practices. In: Ribeiro, RP, Pfahringer, B, Japkowicz, N, Larranaga, P, Jorge, AM, Soares, C, Abreu, PH et al. (eds.) Machine learning and knowledge discovery in databases. Research track, ECML PKDD 2025, part i. Lecture notes in artificial intelligence, vol 16013. Springer, Cham. 265–281. https://doi.org/10.1007/978-3-032-05962-8_16

Ouyang L, Wu J, Jiang X et al. (2022) Training language models to follow instructions with human feedback. In: Advances in Neural Information Processing Systems 35, pp 27730–27744

Parasuraman R, Riley V (1997) Humans and automation: Use, misuse, disuse, abuse. Hum factors 39(2):230–253. https://doi.org/10.1518/001872097778543886

Raile P (2024) The usefulness of ChatGPT for psychotherapists and patients. Humanities Soc Sci Commun 11:47. https://doi.org/10.1057/s41599-023-02567-0

Reber R, Schwarz N (1999) Effects of perceptual fluency on judgments of truth. Conscious cognition 8(3):338–342. https://doi.org/10.1006/ccog.1999.0386

Salvi F, Ribeiro MH, Gallotti R et al. (2025) On the conversational persuasiveness of GPT-4. Nat Hum Behav 9(8):1645–1653. https://doi.org/10.1038/s41562-025-02194-6

Weidinger L, Mellor J, Rauh M et al. (2022) Taxonomy of risks posed by language models. In: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, pp. 214–229. https://doi.org/10.1145/3531146.3533088

Author contributions

GV conceived the comment, developed the conceptual framework, conducted the literature review, and wrote the manuscript. He designed the proposed epistemic auditing framework and approved the final version for submission.

Competing interests

The author declares no competing interests.

Ethics approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Correspondence and requests for materials should be addressed to Giulio Vidotto.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.