Grounding AI in humanistic inquiry

Interdisciplinary challenges for evaluation and interpretability

Oliver Eberle

1. Introduction

The rise of foundation models has enabled machine learning (ML) systems to be applied across a wide range of scientific domains (Richards et al., 2019; Carleo et al., 2019; Rajpurkar et al., 2022; Quazi 2022). Recent advances have further led to the development and rise of artificial large reasoning models (LRMs) (see (Xu et al., 2025) for an overview), which are designed to address multi-step decision tasks like the derivation of mathematical proofs, generating code, and common-sense reasoning. The accelerated use of such LLM systems in scientific practice has especially benefited fields like the natural sciences, characterized by large-scale data, quantitative methods, and well-structured datasets. Meanwhile, HPSS, characterized by low-resource data as well as challenging annotation and evaluation tasks, took a backseat. Nonetheless, AI has the potential to "radically change humanities" (Duch 2023) and offers opportunities to address both technical and societal challenges amid the rapid adoption of this technology.

However, as AI models grow in complexity and are applied to more challenging tasks like artificial reasoning, their inner workings become increasingly opaque, posing challenges for scientific applications that demand interpretability, trustworthiness, and verifiability. Explainable AI and interpretability research aim to develop methods that uncover how AI systems operate at a fundamental level (Samek et al., 2019; Linardatos et al., 2021; Zhao et al., 2024).

The following explores the mutual interactions between ML and HPSS, then discusses the challenges of evaluation and interpretability, before presenting ways forward and concluding.

2. Current interactions between ML and HPSS research

2.1 ML-inspired HPSS

2.1.1 ML accelerates manual HPSS processes

Many works at the intersection of HPSS and LLMs aim to streamline and scale labor-intensive tasks such as transcription, text restoration, entity recognition, topic attribution, and automated image extraction. Nikolaidou et al. (2022) and Sommerschield et al. (2023) provide surveys of relevant data and methods in the historical sciences, highlighting the specific challenges like out-of-domain materials and model evaluation. Similarly, social scientists have used LLMs as a tool for automating manual processes, including annotation or information extraction, as well as summarizing and analyzing large amounts of speech and text (see Macanovic 2022; Linegar et al., 2023 for an overview). For a broader review of ML and LLM approaches used in the context of the humanities, see Chapinal-Heras and Díaz-Sánchez (2023), Simons et al. (2026), Meding and Daugs (2026), and Lang (2026).

2.1.2 ML can promote novel HPSS research

The iterative process of gathering, analyzing, and interpreting material in HPSS research challenges standard prediction systems' ability to produce novel insights. Consequently, relatively few works have employed ML in the context of historical insight discovery and "historian in the loop" studies (Assael and Sommerschield et al. 2022; Eberle et al. 2024; Assael and Sommerschield et al., 2025). Further advances would require ML and LLM systems to deeply understand relevant sources, including historical foundation models evaluated on expert global history knowledge, as recently explored by Hauser et al. (2024). Social scientists have used LLMs to study the modeling of cultural transmission (Brinkmann et al., 2023, Lu et al., 2025), the emergence of psychological theories in LLMs, e.g., theory of mind (Strachan et al., 2024) or behavioural game theory (Akata et al., 2025), and their usefulness for LLM social simulations (Qu and Wang 2024; Anthis et al., 2025). This demonstrates how advances in ML can support and enable the investigation of novel scientific questions in HPSS.

2.1.3 ML as an application of the HPSS

Philosophy has interacted with ML by challenging its concepts and the integration of AI progress into existing theories, e.g., bringing statistical learning theory into discussions on the reliability of induction (Harman and Kulkarni 2007), viewing ML as an "experimental philosophy of science" (Korb 2004), bridging philosophical accounts of causality in ML and philosophy (Pearl 2000; Halpern and Pearl 2005), or how information theory contributed to shaping theories of mind and meaning (Adams 2003). Further exchange of ML and philosophy has centered on concepts of *scientific explanation* (Woodward and Ross 2003) aimed toward interpreting today's largely opaque models as discussed in (Miller 2018; Erasmus et al., 2021). Philosophical analyses of ML systems (Thagard, 1990) anticipated challenges that would become practically relevant to deep learning research decades later, including adversarial or counterfactual data and models that generate predictions based on irrelevant inferences in specific problem-solving scenarios (Szegedy et al., 2014; Lapuschkin et al., 2019). Together, these perspectives position HPSS as an important intellectual framework capable of shaping ML research by driving deeper inquiry into the nature of knowledge, explanation, and understanding.

2.2 HPSS-inspired ML

2.2.1 HPSS challenges ML

Heterogeneity alongside temporality and context-awareness present key challenges to ML and LLMs, which currently limit the faithful processing of historical sources. To address data heterogeneity of complex materials, custom ML approaches have been proposed (e.g., Assael and Sommerschield et al., 2022; Eberle et al., 2024), currently requiring tailored solutions and active model development. Nonetheless, advances in AI and the increasing accessibility of powerful foundation models may soon make AI-based analyses feasible, even in challenging scenarios. For instance, deep optical character recognition (OCR) models have recently been proposed to compress historical long-context data into visual tokens rather than text tokens (Wei et al., 2025), capturing rich spatial and contextual information and enabling more efficient indexing, retrieval, and analysis of archival materials, thereby offering a more contextualized understanding of sources.

While making such sources machine-readable is a crucial step toward enabling scientific insight at scale, humanistic inquiry engages with interpretive reasoning, diachronic investigation, and contextual nuance, which currently constitute underexplored areas of ML research. For instance, the computational modeling of temporal change presents a current frontier as discussed in (Büttner, 2026). Given the methodological challenges of extending ML modeling frameworks, recent studies have productively focused on decomposing historical workflows (e.g., Sommerschield et al., 2023) to enable computational analysis, providing valuable intermediate steps toward the development of more sophisticated modeling approaches.

As LLMs are trained to capture statistical associations, they struggle in HPSS domains where causal structures are shaped by complex contexts, e.g., historical, cultural, or institutional, and where scientific explanations depend as much on interpretation as on observation. For example, philosophical inquiry and reasoning may not deliver sufficiently many recognizable, repeatable patterns to robustly train today's AI systems (D'Alessandro, 2026). Philosophers such as (Machamer et al., 2000) have framed scientific practice as uncovering mechanisms that explain how phenomena arise, emphasizing *entities* and how *activities* among them cause change. These ideas were popularized and adopted in biology and neuroscience, which have since informed the analysis of AI systems and, specifically, shaped the concepts and goals of interpretability research as discussed in Section 3.

2.2.2 HPSS grounds ML

While most ML research focuses on technical challenges, HPSS can play a key role in clarifying trade-offs between technological advances and societal values, for example by discussing how research that involves LLMs intersects with issues of data privacy, transparency, and fairness. Epistemology can further ground ML research by examining how evidence and uncertainty constrain what can be confidently known (Lipton, 2004; Kiureghian and Ditlevsen, 2009), providing important conceptual foundations for developing epistemologically informed ML. In this context, research on artificial reasoning with LRMs has mostly focused on deductive settings, e.g., generating mathematical proofs, for which clear ground truth reasoning steps can be obtained for validation. Addressing uncertainty and non-deductive reasoning remains a significant challenge, one in which the HPSS community has long-standing expertise that can aid in grounding future developments. Recent work on analyzing LLMs has further explored neural network representations related to truthfulness (Azaria and Mitchell, 2023) and morality (Wynn et al., 2024; Jiang et al., 2025), building on key philosophical notions and recent benchmarks like MoralBench (Ji et al., 2024). Further empirical evidence of *universal representations* across foundation models and tasks has sparked discussions on converging representations in deep learning (Huh et al., 2024), which the authors conceptually link to Plato's *Allegory of the Cave* and *convergent realism* (Laudan, 1981; Putnam, 1982).

2.2.3 HPSS can inform the development of ML methods

Besides challenging ML, HPSS offers rich data, complex task settings and workflows, as well as methodological guidance for improving model reasoning and alignment. Model tuning through HPSS materials has recently been explored; e.g., moral alignment via in-context ethical policies (Rao et al., 2023), factuality-aware learning algorithms (Lin et al., 2024), and even fine-tuning LLMs on the works of philosopher Daniel Dennett (Schwitzgebel et al., 2024). In the context of designing AI agent systems, philosophical and social science frameworks have informed recent work, including multi-agent systems grounded in epistemology (Shoham and Leyton-Brown, 2008), LLM-agentic philosophers (Barkol, 2025), and cooperative systems (Ashery et al., 2025). Beyond guiding model design, critically evaluating and understanding these systems from perspectives such as epistemology, reasoning, human values, and historical and social context is key to developing AI that is interpretable, reliable, and socially aligned.

3. The challenge of LLM evaluation and interpretability

3.1 Behavioral evaluation of performance

The evaluation of today's state-of-the-art LLMs predominantly focuses on task suites designed to test certain desired capabilities such as general, potentially multilingual, language understanding and comprehension, code generation, mathematical and logical reasoning, and general prompt instruction following; see (Srivastava et al., 2023) for an overview of commonly used tasks and task structure. Aside from evaluating models in terms of their alignment with general human preferences and their performance relative to weakly defined safety and ethics-critical standards, dedicated HPSS benchmarks are largely missing. As standard ML evaluations are limited in capturing the complexity of large text as produced by LLMs, Wallach et al. (2025) propose to frame generative AI evaluation as a social science measurement challenge, providing a path forward in capturing abstract concepts such as ideology, democracy, or bias via measurement theory frameworks. This highlights the potential of HPSS approaches to develop nuanced evaluations of LLM behavior, moving beyond the simplistic technical benchmarks that dominate ML research.

3.2 Understanding AI models through explanation techniques and interpretability studies

While accurate performance measurements are an important step toward using LLMs in HPSS, uncovering and verifying their underlying inference mechanisms are crucial, especially in the context of grounding scientific discovery. Although models may be able to generate novel outputs and scientifically valuable hypotheses without revealing their internal processes, understanding these mechanisms can enhance trust and provide deeper insight. This requires access to faithful model explanations that aim to provide (i) transparency and trustworthiness, (ii) verifiability, (iii) human interpretability and understanding, (iv) insights into potentially unexpected model strategies, and (v) actionable guidance for model improvement or error mitigation. Since explanations can be derived at multiple levels and provide insight in various formats, the following summarizes the most common variants in the context of language models.

3.3 Explanations across levels and structures

Initial explanation efforts have focused on explanations in terms of relevant input features. Such feature attribution methods have been widely used to compute heatmaps over input samples such as token sequences used in text classification tasks (Bach et al., 2015; Ribeiro et al.; 2016, Sundararajan et al., 2017; Ali et al., 2022). Furthermore, the analysis of model internals, i.e., neuron-level interpretation, layer-wise representations and analysis of relevant model components like attention heads, can provide deeper insight into how the model inputs are represented, interpreted and contextualized, aiming to reveal mechanistic strategies. Mechanistic interpretability herein developed methods for the extraction of circuits via causal abstractions of LLMs (Geiger et al., 2021; Geiger et al., 2025). Structured explanations extend beyond heatmaps by capturing higher-order interactions among multiple features (Eberle et al., 2022; Schnake et al., 2022), for example revealing token–token interactions in LLMs (Vasileiou and Eberle, 2024). Algorithmic explanations further aim to identify the implemented internal algorithms of LLMs and LRMs, focusing on defining primitives and compositions thereof (Eberle et al., 2025; Lippl et al., 2025). To approximate the function of specific internal features, such as individual neurons or groups of neurons, feature description methods provide text-based concept summaries that characterize the patterns most strongly associated with their activation patterns (Singh et al., 2023; Bills et al., 2023; Kopf et al., 2025).

Extracting influential dataset samples provides explanations in reference to the training corpus, for example by revealing which samples strongly influenced the learning of a particular pattern (Han et al., 2020; Grosse et al., 2023). Complementary work leverages LLMs' generative abilities to communicate explanations directly, adopting an output-centric view. Methods include *self-explanations* and *chain-of-thought* (Wei et al., 2022; Huang et al., 2023), as well as free-form interactive user prompt explanations such as “explain to me,” “provide evidence,” “what would happen if…”(Slack et al., 2023), and counterfactual explanations (Chen et al., 2024).

3.4 Evaluating the quality of explanations

Across all these variants, explainable AI methods need to be rigorously evaluated to ensure providing reliable insight into the model. Evaluating explanations is challenging due to the wide range of methods and explanatory levels, the lack of agreement on what makes a good explanation, and the inherently fuzzy nature of the concept itself. Common evaluation criteria include *fidelity* to the model’s computations, *plausibility* for human users, and *causality*, which assesses whether explanations capture genuine cause and effect relationships (Samek et al., 2019; Zhao et al., 2024). Evaluation is further complicated by the absence of ground truth explanations and the inherently subjective nature of what human users consider an effective explanation. See also (Zhou et al., 2021) for a review on evaluating explanations, and (Zhao et al., 2024) for LLM evaluations specifically.

In the context of HPSS research, evaluating model explanations is challenging because clearly defined task settings and ground truth annotations are scarce. Advances in LLM-based artificial reasoning and multi-agent collaboration now make it possible to tackle increasingly complex tasks, including AI-driven logic problem solving, commonsense reasoning, and behavioral simulations, and integration into the scientific process (Zhang et al., 2025). This, in turn, calls for datasets and benchmarks that, for example, could focus on the AI-based reproduction of traditional HPSS research findings like the re-discovery of scientific concepts and ideas from predefined sources. Such investigations highlight the importance of expertise across HPSS fields, for example in analyzing a model’s epistemological justification process or studying the collective behavior and collaboration of multiple AI agents. Achieving this will require closer collaboration and ongoing exchange of knowledge, data, and tools between HPSS and AI research.

4. Call for action

4.1 Advancing materials and benchmarks for training and evaluation

HPSS communities must lead the development of domain-specific, curated datasets that reflect their complexities, structure, and contextual richness. Such datasets are essential not only for training more reliable and representative models but also for developing evaluation benchmarks that move beyond generic and narrowly focused metrics. Current benchmarks lack key features of HPSS problems, such as temporal change and cultural grounding, navigating ambiguity and normative value judgments, and working with interpretative or incomplete evidence; see (Hershcovich et al., 2022; Ziems et al., 2024; Morehouse et al., 2025) for some first steps in this direction. These challenges further motivate the development of new evaluation methods that allow capturing dynamic sources or ill-defined labels.

4.2 Grounding LLM and interpretability research in HPSS frameworks

Interpretability and evaluation methods should be explicitly grounded in HPSS conceptual frameworks, incorporating epistemological rigor, historical context, and cultural sensitivity. This will ensure that models are not only technically sound but also socially and scientifically meaningful. The rise of AI has already and will continue to change how we access, process and retrieve information. However, as Offert (2024) argues, traditional humanist frameworks may not fully suffice to conceptualize history in a world of constantly remediated data, challenging HPSS to adapt and extend its frameworks in response to advances in AI. For example, Willems et al. (2025) highlight the need for philosophy in explainable AI and interpretability research to clarify definitions and drive methodological advances for assessing societally critical aspects.

4.3 Participation in shaping AI discussions and development

HPSS scholars must actively contribute to AI ethics, policy, and public discourse, ensuring that debates about LLM development and deployment are informed by deep understanding of historical, philosophical and social implications. Critical perspectives have played a key role in shaping discussions and raising awareness of fairness, biased predictions, and accountability in learning-based systems (Dwork et al., 2012; Shah et al., 2020; Gallegos et al., 2024), and the recent rise of AI agents has further prompted calls for new ethical frameworks (Gabriel et al., 2025). Furthermore, engagement in AI research communities, conferences, and collaborative projects is needed to shape future research agendas, standards, and methodologies that reflect the needs and values of their fields. The organization of joint workshops like the NeurIPS 2011 workshop on *Philosophy and Machine Learning*, the *Machine Learning for Ancient Languages* Series at ACL 2024 and 2025, and the NeurIPS 2025 workshop on *Algorithmic Collective Action* herein provide important opportunities for extending interdisciplinary exchange.

5. Conclusion

HPSS has the unique opportunity and responsibility to shape the future development of AI by providing conceptual frameworks, experience with historically grounded context, and culturally aware evaluation methods. By defining its own benchmarks, datasets, and interpretability standards, HPSS can support the development of AI systems that are not only technically proficient but demonstrate epistemic robustness, ethical integrity, and regulatory compliance. This requires sustained collaboration between HPSS and ML research communities to bridge conceptual gaps and inform methodologies. Such engagement will guide the creation of AI systems that advance human understanding while remaining accountable to the diverse values and complexities of the societies they serve.[1]

1 This chapter was reviewed with support of large language models (LLMs) and its use was restricted to improving grammar and style. The author remains fully responsible for the final version. For details on the use of LLMs in this volume, see the statement in the volume's introduction.

References

Adams F (2003) The informational turn in philosophy. Minds and Machines 13(4):471–501.

Akata E, Schulz L, Coda-Forno J, et al. (2025) Playing repeated games with large language models. Nat Hum Behav 9, 1380–1390. https://doi.org/10.1038/s41562-025-02172-y

Ali A, Schnake T, Eberle O, et al. (2022) XAI for transformers: Better explanations through conservative propagation. In International conference on machine learning (pp. 435–451). PMLR.

Anthis JR, Liu R, Richardson SM, et al. (2025) Position: LLM Social Simulations Are a Promising Research Method. In Forty-second International Conference on Machine Learning Position Paper Track.

Ashery AF, Aiello LM and Baronchelli A (2025) Emergent social conventions and collective bias in LLM populations. Science Advances, 11(20), eadu9368.

Assael Y, Sommerschield T, Cooley A, et al. (2025) Contextualizing ancient texts with generative neural networks. Nature. https://doi.org/10.1038/s41586-025-09292-5

Assael Y, Sommerschield T, Shillingford B, et al. (2022) Restoring and attributing ancient texts using deep neural networks. Nature 603, 280–283. https://doi.org/10.1038/s41586-022-04448-z

Azaria A and Mitchell T (2023) The Internal State of an LLM Knows When It's Lying. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 967–976, Singapore. Association for Computational Linguistics.

Bach S, Binder A, Montavon G, et al. (2015) On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLoS ONE 10(7): e0130140. https://doi.org/10.1371/journal.pone.0130140

Barkol D (2025) Agentic AI Explained: A Philosophical Framework for Understanding AI Agents. Microsoft Dev Blogs. https://devblogs.microsoft.com/all-things-azure/agentic-philosophers/.

Bills S, Cammarata N, Mossing M, et al. (2023) Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html.

Büttner J (2026) Why pursue temporally-grounded AI for historical disciplines, and what makes it so challenging? In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-2.

Brinkmann L, Baumann F, Bonnefon JF et al. (2023) Machine culture. Nat Hum Behav 7, 1855–1868. https://doi.org/10.1038/s41562-023-01742-2

Carleo G, Cirac I, Cranmer K, et al. (2019) Machine learning and the physical sciences. Reviews of Modern Physics, 91(4), 045002.

Chapinal-Heras D and Díaz-Sánchez C (2023) A review of AI applications in Human Sciences research. Digital Applications in Archaeology and Cultural Heritage, 30, e00288.

Chen Y, Zhong R, Ri N, et al. (2024) Do models explain themselves? counterfactual simulatability of natural language explanations. In Proceedings of the 41st International Conference on Machine Learning (pp. 7880–7904).

D'Alessandro W (2026) LLMs as philosophers: what can they do, and why aren't they better? In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-1.

Der Kiureghian A and Ditlevsen O (2009) Aleatory or epistemic? Does it matter?. Structural safety, 31(2), 105–112.

Duch W (2024) Artificial Intelligence and the Limits of the Humanities. Er (r) go. Teoria-Literatura-Kultura, (48), 269–297.

Dwork C, Hardt M, Pitassi T, et al. (2012) Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference (pp. 214–226).

Eberle O, Büttner J, Kräutli F, et al. (2022) Building and interpreting deep similarity models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(3):1149–1161. doi: 10.1109/TPAMI.2020.3020738.

Eberle O, McGee T, Giaffar H, et al. (2025) Position: We Need An Algorithmic Understanding of Generative AI. In Forty-second International Conference on Machine Learning Position Paper Track.

Erasmus A, Brunet TDP and Fisher E (2021) What is Interpretability?. Philos. Technol. 34, 833–862. https://doi.org/10.1007/s13347-020-00435-2

Fabian O (2024) On the Concept of History (in Foundation Models). Forthcoming in: Thinking with AI, ed. Hannes Bajohr, Open Humanities Press.

Gabriel I, Keeling G, Manzini A, et al. (2025) We need a new ethics for a world of AI agents. Nature. Aug;644(8075):38-40. doi: 10.1038/d41586-025-02454-5. PMID: 40764692.

Gallegos IO, Rossi RA, Barrow J, et al. (2024) Bias and Fairness in Large Language Models: A Survey. Computational Linguistics; 50 (3): 1097–1179. DOI: https://doi.org/10.1162/coli_a_00524

Geiger A, Ibeling D, Zur A, et al. (2025) Causal abstraction: A theoretical foundation for mechanistic interpretability. Journal of Machine Learning Research, 26(83), 1–64.

Geiger A, Lu H, Icard T, et al. (2021) Causal abstractions of neural networks. Advances in Neural Information Processing Systems, 34, 9574–9586.

Grosse R, Bae J, Anil C, et al. (2023) Studying large language model generalization with influence functions. arXiv preprint arXiv:2308.03296.

Halpern J and Pearl J (2005) Causes and explanations: A structural-model approach. British J. Phil. Sci. 56:843-911.

Han X, Wallace BC, and Tsvetkov Y (2020) Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5553–5563, Online. Association for Computational Linguistics.

Harman G and Kulkarni S (2007) Reliable Reasoning: Induction and Statistical Learning Theory. MIT Press, Cambridge, MA.

Hauser J, et al. (2024) Large language models' expert-level global history knowledge benchmark (HiST-LLM). Advances in Neural Information Processing Systems 37: 32336–32369.

Hershcovich D, Frank S, Lent H, et al. (2022) Challenges and Strategies in Cross-Cultural NLP. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 6997–7013).

Huang J, Gu S, Hou L, et al. (2023) Large Language Models Can Self-Improve. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 1051–1068, Singapore. Association for Computational Linguistics.

Huh M, Cheung B, Wang T, et al. (2024) Position: The platonic representation hypothesis. In Forty-first International Conference on Machine Learning.

Ji J, Chen Y, Jin M, et al. (2025) Moralbench: Moral evaluation of llms. ACM SIGKDD Explorations Newsletter, 27(1), 62–71.

Jiang L, Hwang JD, Bhagavatula C, et al. (2025) Investigating machine moral judgement through the Delphi experiment. Nat Mach Intell 7, 145–160. https://doi.org/10.1038/s42256-024-00969-6

Kopf L, Feldhus N, Bykov K, et al. (2025) Capturing Polysemanticity with PRISM: A Multi-Concept Feature Description Framework. Advances in neural information processing systems, 39.

Korb K (2004) Introduction: Machine learning as philosophy of science. Minds and Machines 14(4).

Lang S (2026) Critical concerns for using LLMs in the (computational) humanities and beyond. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-1.

Lapuschkin S, Wäldchen S, Binder A, et al. (2019) Unmasking Clever Hans predictors and assessing what machines really learn. Nat Commun 10, 1096. https://doi.org/10.1038/s41467-019-08987-4

Laudan L A (1981) Confutation of Convergent Realism. Philosophy of Science; 48(1):19-49. DOI:10.1086/288975

Lazer D, Pentland A, Adamic L, et al. (2009) Social science. Computational social science. Science. Feb 6;323(5915):721-3. doi: 10.1126/science.1167742. PMID: 19197046; PMCID: PMC2745217.

Lin SC, Gao L, Oguz B, et al. (2024) Flame: Factuality-aware alignment for large language models. Advances in Neural Information Processing Systems, 37, 115588–115614.

Linardatos P, Papastefanopoulos V and Kotsiantis S (2021) Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy, 23(1), 18. https://doi.org/10.3390/e23010018

Linegar M, Kocielnik R and Alvarez RM (2023) Large language models and political science. Front. Polit. Sci. 5:1257092. doi: 10.3389/fpos.2023.1257092

Lippl S, McGee T, Lopez K, et al. (2025) Algorithmic Primitives and Compositional Geometry of Reasoning in Language Models, arXiv preprint arXiv:2510.15987.

Lipton P (2004) Induction. Chapter 1 of Inference to the Best Explanation.

Lu JG, Song LL and Zhang LD (2025) Cultural tendencies in generative AI. Nat Hum Behav. https://doi.org/10.1038/s41562-025-02242-1

Macanovic A (2022) Text mining for social science – The state and the future of computational text analysis in sociology. Soc Sci Res. Nov;108:102784. doi: 10.1016/j.ssresearch.2022.102784. Epub 2022 Sep 2. PMID: 36334929.

Machamer P, Darden L, Craver CF (2000) Thinking about Mechanisms. Philosophy of Science; 67(1):1-25. doi:10.1086/392759

Meding H and Daugs A (2026) On the use and limitations of large language models in historical scholarship. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-1.

Miller T (2018) Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1–38.

Morehouse K, Swaroop S and Pan W (2025) Position: Rethinking LLM Bias Probing Using Lessons from the Social Sciences. Forty-second International Conference on Machine Learning Position Paper Track.

Nikolaidou K, Seuret M, Mokayed H, et al. (2022) A survey of historical document image datasets. IJDAR 25, 305–338. https://doi.org/10.1007/s10032-022-00405-8

Pearl J (2000) Causality: Models, Reasoning, and Inference. Cambridge: Cambridge University Press.

Putnam H (1982) Three Kinds of Scientific Realism. The Philosophical Quarterly (1950-) 32 (128): 195–200.

Qu Y and Wang J (2024) Performance and biases of Large Language Models in public opinion simulation. Humanit Soc Sci Commun 11, 1095. https://doi.org/10.1057/s41599-024-03609-x

Quazi S (2022) Artificial intelligence and machine learning in precision and genomic medicine. Medical Oncology, 39(8), 120.

Rajpurkar P, Chen E, Banerjee O, et al. (2022) AI in health and medicine. Nat Med 28, 31–38. https://doi.org/10.1038/s41591-021-01614-0

Rao AS, Khandelwal A, Tanmay K, et al. (2023) Ethical Reasoning over Moral Alignment: A Case and Framework for In-Context Ethical Policies in LLMs. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 13370–13388, Singapore. Association for Computational Linguistics.

Ribeiro MT, Singh S and Guestrin C (2016) "Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135–1144).

Richards BA, Lillicrap TP, Beaudoin P, et al. (2019) A deep learning framework for neuroscience. Nat Neurosci 22, 1761–1770. https://doi.org/10.1038/s41593-019-0520-2

Samek W, Montavon G, Vedaldi A, et al. (2019) Explainable AI: interpreting, explaining and visualizing deep learning (Vol. 11700). Springer Nature.

Schnake T, Eberle O, Lederer J, et al. (2022) Higher-order explanations of graph neural networks via relevant walks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(11):7581–7596. doi: 10.1109/TPAMI. 2021.3115452.

Schwitzgebel E, Schwitzgebel D and Strasser A (2024) Creating a large language model of a philosopher. Mind & Language, 39(2), 237–259.

Shah DS, Schwartz HA and Hovy D (2020) Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5248–5264, Online. Association for Computational Linguistics.

Shoham Y and Leyton-Brown K (2008) Multiagent systems: Algorithmic, game-theoretic, and logical foundations. Cambridge University Press.

Simons A, Zichert M and Wüthrich A (2026) Large language models for history, philosophy, and sociology of science: Interpretive uses, methodological challenges, and critical perspectives. *Studies in History and Philosophy of Science* 117: 102151. https://doi.org/10.1016/j.shpsa.2026.102151.

Singh C, Hsu AR, Antonello R, et al. (2023) Explaining black box text modules in natural language with language models. arXiv preprint arXiv:2305.09863.

Slack D, Krishna S, Lakkaraju H, et al. (2023) Explaining machine learning models with interactive natural language conversations using TalkToModel. Nat Mach Intell 5, 873–883. https://doi.org/10.1038/s42256-023-00692-8

Sommerschield T, Assael Y, Pavlopoulos J, et al. (2023) Machine Learning for Ancient Languages: A Survey. Computational Linguistics; 49 (3): 703–747. doi: https://doi.org/10.1162/coli_a_00481

Srivastava A, Rastogi A, Rao A, et al. (2023) Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Transactions on machine learning research.

Strachan JWA, Albergo D, Borghini G. et al. (2024) Testing theory of mind in large language models and humans. Nat Hum Behav 8, 1285–1295. https://doi.org/10.1038/s41562-024-01882-z

Sundararajan M, Taly A and Yan Q (2017) Axiomatic attribution for deep networks. In International conference on machine learning (pp. 3319–3328). PMLR.

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow IJ and Fergus R (2014) Intriguing properties of neural networks. ICLR (Poster)

Thagard P (1988) Computational Philosophy of Science. MIT Press, Cambridge, MA.

Thagard, Paul (1990) Philosophy and Machine Learning. Canadian Journal of Philosophy 20, no. 2: 261–76. http://www.jstor.org/stable/40231695.

Vasileiou A and Eberle O (2024) Explaining Text Similarity in Transformer Models. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7859–7873, Mexico City, Mexico. Association for Computational Linguistics.

Wallach H, Desai M, Cooper AF, et al. (2024) Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge. In Forty-second International Conference on Machine Learning Position Paper Track.

Wei H, Sun Y and Li Y (2025) DeepSeek-OCR: Contexts Optical Compression, arXiv preprint arXiv:2510.18234.

Wei J, Wang X, Schuurmans D, et al. (2022) Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824–24837.

Woodward J and Ross L (2003) Scientific Explanation. In E. N. Zalta (Ed.), The Stanford encyclopedia of philosophy. Metaphysics Research Lab, Stanford University.

Wynn AH, Sucholutsky I and Griffiths TL (2024) Learning human-like representations to enable learning human values. Advances in Neural Information Processing Systems, 37, 30230–30260.

Xu F, Hao Q, Zong Z, et al. (2025) Towards large reasoning models: A survey of reinforced reasoning with large language models. Patterns, Volume 6, Issue 10, 2025, https://doi.org/10.1016/j.patter.2025.101370.

Zhang Y, Khan SA, Mahmud A, et al. (2025) Exploring the role of large language models in the scientific method: from hypothesis to discovery, npj Artificial Intelligence, 1, p.14. doi: 10.1038/s44387-025-00019-5.

Zhao H, Chen H, Yang F, et al. (2024) Explainability for large language models: A survey. ACM Transactions on Intelligent Systems and Technology, 15(2), 1–38.

Zhou J, Gandomi AH, Chen F, et al. (2021) Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics, 10(5), 593.

Ziems C, Held W, Shaikh O, et al. (2024) Can Large Language Models Transform Computational Social Science? Computational Linguistics, 50(1):237–291.