The Missing Layer in Modern AI: Interpretation

A Narrative Conceptual Review

Michael Daniel Negrea — Chief Executive Officer @ Genosen — michael@genosen.ai

Abstract

Modern artificial intelligence (AI) systems have achieved unprecedented performance on tasks of prediction and generation. A recurring concern across multiple disciplines, however, is that such systems insufficiently model the receiver-side process by which a human assigns meaning to an output. This paper develops the position that interpretation, defined narrowly as the cognitive-semiotic process through which a human recipient constructs perceived meaning from a stimulus in context, constitutes a distinct, addressable layer of AI-mediated communication that current paradigms address only obliquely. The paper proceeds as a narrative conceptual review. It synthesises peer-reviewed literature from artificial intelligence, human–computer interaction (HCI), cognitive science, behavioural economics, communication theory, and information systems; engages with counterexamples drawn from reinforcement learning from human feedback (RLHF), user modelling, recommender systems, human–AI interaction, and adaptive interfaces; and argues for treating interpretation as a research object in its own right, alongside the existing concerns of interpretability, understanding, explanation, comprehension, and trust. The review's contribution is conceptual and threefold: a synthesis across literatures that rarely converse with one another, a critique of the proxy assumptions on which contemporary evaluation regimes rest, and a research agenda. The argument is hedged: the reviewed literatures do not identify a single unanimous omission but a family of complementary gaps that together motivate the case for an interpretive layer.

1. Introduction

Predictive and generative capacities have advanced rapidly in the past decade. Large neural language and vision models now produce fluent text, photorealistic imagery, and accurate forecasts across an expanding range of tasks (Bender et al., 2021; Mitchell & Krakauer, 2023). The trajectory has revived an older and

unresolved question: do these systems understand what they process, or do they predict what is likely to follow (Bender & Koller, 2020; Searle, 1980)? The question is not merely philosophical. As AI is embedded in decision environments (from health information delivery to consumer-facing interfaces), the gap between what a system generates and what a human ultimately perceives, comprehends, and acts upon becomes a substantive design and ethical concern (Lee & See, 2004; Shneiderman, 2020).

This paper takes the claim that modern AI is highly capable at prediction and generation but insufficiently models human interpretation, and treats it as a hypothesis to be examined against the existing peer-reviewed literature. Rather than proposing a new architecture or implementation, the paper synthesises theoretical and empirical work across disciplines, defines its terms with care, and engages with counterexamples that complicate the headline claim.

1.1 Defining interpretation

The argument hinges on a specific definition. Interpretation, as used in this paper, refers to the receiver-side cognitive-semiotic process through which a human, situated in a particular context, constructs a perceived meaning from an external stimulus. It is a process, not a property; it is located in the user, not in the system; and it is context-dependent in the sense that the same stimulus may be interpreted differently by different recipients, by the same recipient at different times, or under different framings (Krippendorff, 2019; Tversky & Kahneman, 1981). Interpretation thus encompasses the perceptual encoding of the stimulus, the activation of relevant prior knowledge, inferential elaboration including framing and causal attribution, and the formation of a working semantic representation on which downstream judgement, trust, and action are based (Kahneman, 2011; Petty & Cacioppo, 1986).

Five distinctions are heuristic but sufficient to mark interpretation as a distinct object of inquiry.

- **Interpretability** is a property of a model: the degree to which its mechanisms or outputs are intelligible to a human observer (Doshi-Velez & Kim, 2017; Lipton, 2018; Rudin, 2019). Interpretability is system-side; interpretation is user-side. A model can be interpretable in principle while still being misinterpreted in practice.
- **Understanding** denotes a deeper grasp of meaning, causal structure, or world-modelling, applied to either system or user (Lake et al., 2017; Mitchell & Krakauer, 2023; Pearl, 2019; Searle, 1980).

Understanding can be present without explicit interpretation, and interpretation can occur without genuine understanding.

- **Explanation** is a communicative act: an account given by one party to another (Miller, 2019). Explanations are inputs to interpretation rather than substitutes for it.
- **Comprehension** denotes successful decoding of intended content, in the sense that interpretation aligns with the sender's intent (Sweller, 1994). Comprehension is a particular outcome of interpretation; misinterpretation is also a form of interpretation.
- **Trust** is an attitudinal stance about the reliability or benevolence of a system, often formed under uncertainty and used to guide reliance (Lee & See, 2004; Mayer, Davis, & Schoorman, 1995). Trust is downstream of interpretation: a user's interpretation of system outputs shapes the trust they extend.

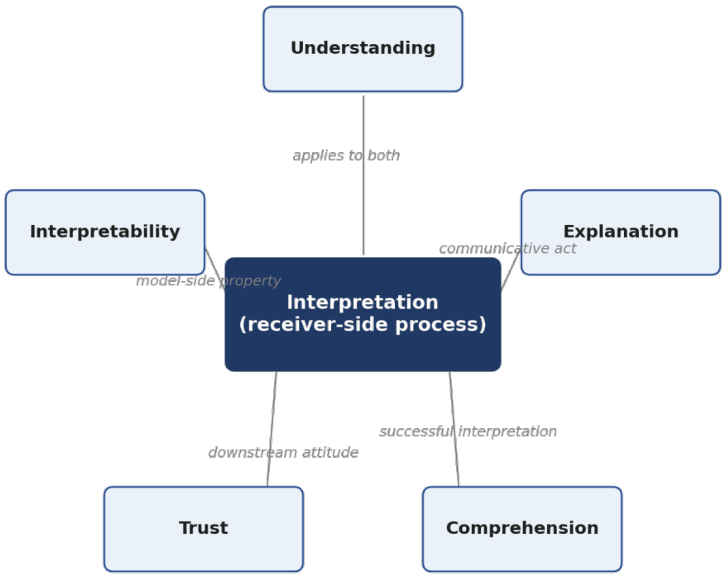

Figure 1. Interpretation in relation to adjacent concepts. Interpretation is positioned as the receiver-side process, distinct from but linked to interpretability, understanding, explanation, comprehension, and trust.

1.2 Method

This paper is a narrative conceptual review rather than a systematic one. It does not claim exhaustive coverage of any single literature, nor does it follow PRISMA or comparable protocols. Its purpose is

conceptual synthesis across literatures that have evolved largely in parallel. Sources were selected to reflect foundational and highly cited works in each relevant discipline, recent peer-reviewed contributions articulating the debate around understanding and human-centred AI, and works bearing on counterexamples to the headline claim. Priority was given to journal articles, peer-reviewed conference proceedings (notably ACL, CHI, FAccT, UIST, AAAI, NeurIPS), and academic books from established publishers. A small number of widely cited preprints (e.g., Doshi-Velez & Kim, 2017) are included where they have become canonical references in the peer-reviewed literature that subsequently cites them. Three limitations apply: narrative reviews carry inherent selection bias; breadth across disciplines comes at the cost of depth within each; and the literatures synthesised here are evolving rapidly. The position advanced is offered as a working conceptual frame rather than a final account.

1.3 Contribution

This paper does not present new empirical results, propose a novel architecture, or disclose implementation details. Its contribution is conceptual and threefold. First, it assembles, from literatures that rarely converse with one another, a set of complementary claims that together motivate treating interpretation as a distinct layer of AI-mediated communication. Second, it questions the proxy assumptions on which much contemporary evaluation rests, in particular the use of behavioural engagement metrics, fluency benchmarks, and preference rankings as substitutes for cognitive and interpretive states. The critique is hedged: these proxies are not without value, but they are routinely treated as more direct measures of meaning, comprehension, and trust than the empirical literature supports. Third, it sketches a conceptual direction for future research without proposing any specific architecture or pipeline. The paper is therefore a position paper in the sense of Doshi-Velez and Kim (2017) and Bender and Koller (2020): a structured argument that aims to influence how a problem is framed, not a study that aims to settle it.

2. Modern AI's Predictive Paradigm

A persistent line of critique across philosophy of mind, cognitive science, and machine learning distinguishes prediction from understanding. Searle (1980), through the Chinese Room thought experiment, argued that syntactic symbol manipulation (however behaviourally competent) does not entail semantic

understanding. The argument remains contested but continues to frame contemporary debates about whether statistical systems that pattern-match on form can be said to grasp meaning.

Pearl (2019) provides a computational reformulation: a three-rung ladder of causation that separates association (seeing), intervention (doing), and counterfactual reasoning (imagining), with mainstream machine learning operating on the lowest rung. Schölkopf et al. (2021) extend this argument to representation learning, formalising the absence of causal structure as a principal limitation of current models. Lake, Ullman, Tenenbaum, and Gershman (2017) propose that human-like learning requires building causal models that support explanation, grounded in intuitive theories of physics and psychology and harnessing compositionality and learning-to-learn. Bender and Koller (2020) make a complementary argument from a linguistic standpoint: a system trained only on linguistic form has, a priori, no path to meaning, because meaning is constituted by the relation between form and communicative intent in the world. Mitchell and Krakauer (2023) frame the contemporary debate around large language models as one about the very definition of understanding, noting that surface fluency does not guarantee robust abstraction or world-modelling.

The convergence across these works supports a conceptual distinction between predictive accuracy on benchmarks and the structures associated with genuine understanding, even where the empirical question of whether further scale can close the gap remains open. The figure below renders the distinction in its most stripped form.

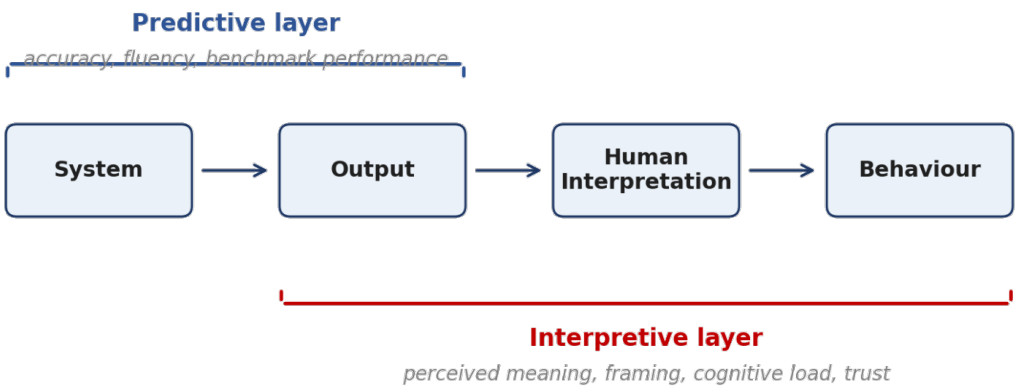

Figure 2. The two layers of AI-mediated communication. Predictive performance and interpretive uptake are conceptually separable links in the chain through which AI outputs become consequential.

The point of the figure is that predictive performance and interpretive uptake are conceptually separable links in the chain through which AI outputs become consequential. A system can perform well on the first link and poorly on the second; equally, the design effort spent on the first does not automatically transfer to the second.

3. Human Interpretation as a Missing Layer

Interpretation is constructive rather than passive. Krippendorff (2019), in his canonical treatment of content analysis, frames communication data as artefacts "created and disseminated to be seen, read, interpreted, enacted, and reflected upon according to the meanings they have for their recipients," explicitly rejecting the container metaphor in which meaning is treated as a property of the message rather than of the interpretive relation. Tversky and Kahneman (1981) demonstrate empirically that logically equivalent presentations of the same information produce systematically different preferences, illustrating that perceived meaning is partly constituted by linguistic and contextual presentation rather than by the underlying state of the world.

In human–computer interaction, Norman (2013) describes how users construct mental models of systems from visible signifiers, feedback, affordances, and prior experience, and how usability failures often arise when the designer's conceptual model and the user's mental model diverge. Hassenzahl and Tractinsky (2006) and Hassenzahl (2010) similarly argue that user experience emerges from the interaction of perception, action, motivation, and cognition with situational context, and is fundamentally subjective and dynamic. Across these traditions, what an artefact "means" is the joint product of artefact, context, and the cognitive and emotional state of the receiver.

The conceptual frame below captures the basic move: meaning is located in the recipient and shaped by context, rather than residing in the artefact itself.

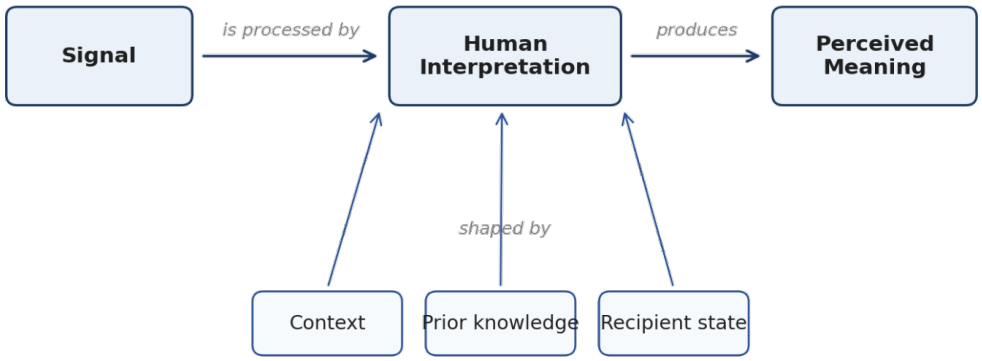

Figure 3. Meaning as a relation rather than a property. Perceived meaning is the joint product of signal, recipient state, prior knowledge, and context.

The implication for AI research is that engagement with the receiver side cannot be deferred to downstream usability work. The construction of perceived meaning is not a usability concern but a structural layer through which the value of any AI output is realised.

4. Why Engagement Is Not Understanding

Behavioural metrics (clicks, dwell time, likes, shares) are widely used as proxies for engagement, and engagement is in turn often treated as a proxy for comprehension, persuasion, or trust. Engagement itself is a polysemic and multidimensional construct: the social-media engagement literature documents at least four broad categories of metrics (quantitative, normalised, indexed, and qualitative) without any single agreed operationalisation. The conceptual slippage from engagement to understanding is therefore unsupported even within the engagement literature itself.

The empirical case for dissociation is stronger. Avram, Micallef, Patil, and Menczer (2020) show experimentally that exposure to social-engagement metrics increases users' vulnerability to low-credibility information, because the metrics are processed as endorsement cues independent of evaluation of the underlying content. Petty and Cacioppo's (1986) elaboration likelihood model provides a theoretical account: central-route processing scrutinises arguments while peripheral-route processing reacts to surface cues including popularity signals, and engagement metrics tend to track peripheral outcomes more reliably than central processing.

The implication is not that behavioural metrics are uninformative. They are. The implication is that they are weak proxies for the interpretive states they are routinely taken to represent. A system optimised only against engagement is optimised against a signal that is dissociable, by construction and by evidence, from the cognitive states it is meant to indicate.

5. Human Cognition in Digital Environments

5.1 Cognitive load

Cognitive Load Theory (Sweller, 1994; Sweller, van Merriënboer, & Paas, 2019) holds that human working memory is severely capacity-limited and that the structure of presented information determines whether comprehension and decision-making proceed efficiently. Intrinsic load is inherent in the material; extraneous load is imposed by presentation; germane load is devoted to schema construction. Extraneous load (for example, when information is fragmented across modalities or redundantly presented) degrades comprehension even when the underlying content is unchanged. The theory has direct implications for AI-generated content and AI-mediated interfaces: complexity that exceeds users' processing capacity produces predictable degradation in comprehension, trust, and decision quality (Norman, 2013). The perceptual and cognitive cost of interpreting a message is not invariant but shaped by design choices.

5.2 Trust

Lee and See (2004), synthesising across organisational, sociological, interpersonal, psychological, and neurological perspectives, show that trust mediates reliance on automation when users cannot fully verify behaviour. Their model identifies performance, process, and purpose as the dimensions on which trust is calibrated, and emphasises that miscalibrated trust (both over-reliance and disuse) is a primary failure mode of automated systems. Mayer, Davis, and Schoorman (1995) provide a parallel organisational model centring on ability, benevolence, and integrity. Across both traditions, trust appears to mediate behavioural readiness: it is the attitudinal state through which interpretation translates into action or inaction.

5.3 Framing and clarity

Tversky and Kahneman (1981) and the broader framing-effects literature (Kahneman, 2011) demonstrate that even small variations in linguistic and presentational framing produce large shifts in interpretation and preference, including in domains as consequential as medical risk and financial choice. The dual-process account formalised by Petty and Cacioppo (1986) explains why: under typical conditions of limited attention and motivation, recipients process messages along the peripheral route, where surface cues dominate. Clarity, understood here as the recoverability of intended meaning under typical conditions of attention, is therefore not a property of a message in isolation but a property of the fit between message, presentation, and audience.

5.4 Sequential and context-dependent processing

A general observation across cognitive psychology, persuasion theory, and HCI is that humans appear to process information sequentially rather than independently, and that earlier states condition later ones. Attention shapes what is encoded; clarity shapes how it is parsed; trust shapes how it is acted upon. The literatures reviewed here treat this conceptually rather than as a unified computational pipeline, and the conceptual observation is sufficient for the present purpose: interpretation unfolds across stages that are not reducible to any single moment.

6. Existing Gaps in AI Systems

When the strands above are read together, they do not articulate a single unanimously identified omission. They identify a family of complementary gaps from different vantage points: causal structure (Pearl, 2019; Schölkopf et al., 2021), intuitive theories and compositional generalisation (Lake et al., 2017), grounded meaning (Bender & Koller, 2020; Bender et al., 2021), social-scientific grounding of explanation (Doshi-Velez & Kim, 2017; Miller, 2019), robust reasoning and common sense (Marcus & Davis, 2019), and abstraction and analogy (Mitchell & Krakauer, 2023). The diagnosis is heterogeneous; the convergence is on the broader observation that the cognitive, communicative, and social conditions of how AI outputs become meaningful for humans are inadequately represented in current architectures.

6.1 Engaging counterexamples

The claim that contemporary AI insufficiently models human interpretation must be tested against several active research programmes that already engage, in different ways, with the receiver side of AI-mediated

communication. These programmes complicate the headline claim rather than refute it; the case for an interpretive layer is strengthened when their limits are made explicit.

Reinforcement learning from human feedback (Christiano et al., 2017; Ouyang et al., 2022) trains models against human preference signals rather than handcrafted reward functions, and is widely credited with making large language models more usable. RLHF undeniably incorporates human signal into training, but it does so under several limiting assumptions. Preferences are collected as pairwise comparisons and aggregated, tending to collapse heterogeneous user populations toward an averaged preference. What is captured is the preferred output, not the cognitive process by which that output is interpreted. Preference signals are themselves products of interpretation by annotators operating under specific instructions, so RLHF inherits any framing or interpretive variability present in the annotation protocol. RLHF aligns generation with human preference outcomes; it does not, in itself, model how a downstream user constructs meaning from a particular output in a particular context.

User modelling (Fischer, 2001; Kobsa, 2001) and contemporary personalisation systems learn user-level parameters from interaction history and adapt outputs accordingly. They typically capture preferences, behavioural patterns, and demographic or contextual features. They do not, in general, model the cognitive processes through which users interpret particular artefacts; they model the artefacts users select. A user model that predicts what a user will click is not equivalent to a model of how the user interprets what is presented.

Recommender systems (Ricci, Rokach, & Shapira, 2015) similarly engage with the user side of communication, but their objective is the prediction of relevance or choice rather than the modelling of interpretation. The recommender-systems literature has acknowledged this gap, with substantial work on trust, explanation, and persuasiveness as factors that mediate the acceptance of recommendations. Even where explanations are offered, evaluation typically remains anchored in behavioural metrics (click-through, acceptance, retention) that are weak proxies for the interpretive state.

Human–AI interaction has produced influential design guidance, including the eighteen guidelines of Amershi et al. (2019) and the human-centred AI framework of Shneiderman (2020) and Shneiderman (2022). Bansal and colleagues (Bansal, Nushi, Kamar, Lasecki, Weld, & Horvitz, 2019; Bansal, Nushi, Kamar, Weld, Lasecki, & Horvitz, 2019) have shown empirically that the alignment between users' mental models and an AI

system's actual error boundaries is a primary determinant of joint team performance, and that updates which improve AI accuracy can degrade team performance when they are incompatible with the user's existing mental model. This work engages directly with the receiver side. It does not undermine the case for an interpretive layer; rather, it supports it. Nonetheless, mental-model alignment remains largely a research instrument rather than a standard, scalable signal available at deployment time.

Adaptive interfaces adjust presentation in response to inferred user state. The most pointed counterexample is the algorithmic-curation literature: Eslami et al. (2015) demonstrated that a majority of Facebook users were unaware that their news feed was algorithmically curated and held inaccurate beliefs about its operation. The study is a paradigm case of the dissociation between system behaviour and user interpretation: the system was personalising effectively on its own terms while users were interpreting the resulting feed as a direct social signal. Adaptive systems operate on the receiver side, but their effects on user interpretation can be substantial and largely invisible to the users themselves, which intensifies rather than resolves the case for explicit attention to the interpretive layer.

These programmes show that contemporary AI is not blind to the receiver. Significant capacities (preference alignment, personalisation, explanation, mental-model-aware design, adaptive presentation) already exist. The argument advanced here is more specific: these capacities engage with proxies for interpretation (preferences, choices, behaviour, accepted explanations) rather than with interpretation itself as defined in Section 1.1. The proxies are useful; they are not equivalent.

6.2 Simulating human reasoning

Park et al. (2023) introduce "generative agents" (language-model-driven simulacra of human behaviour that plan, remember, and interact in simulated environments), illustrating a broader trend of using large language models as proxies for human responses. The limits of such simulation are widely acknowledged. Bender and Koller (2020) and Bender et al. (2021) caution that form-only training does not, in any robust sense, model the speakers whose utterances it reproduces. Mitchell and Krakauer (2023) note that benchmark performance on reasoning tasks is unreliable evidence of underlying competence. From a behavioural-economics perspective, Gigerenzer and Gaissmaier (2011) document that human decision-making relies on fast-and-frugal heuristics adapted to environments of uncertainty; computational approaches that ignore this structure are likely to mispredict human judgement. Future systems may increasingly aim to simulate

heterogeneous populations rather than relying on static personas, but the open conceptual questions about the validity and grounding of such simulation are substantial.

6.3 Key debates

Several genuine debates should be acknowledged. The first is whether understanding is a discrete property or a continuous capacity that can be approached through scale. Bender and Koller (2020) argue from first principles that form-only training cannot yield meaning; emergent-capability arguments (discussed in Mitchell & Krakauer, 2023) hold that the boundary is empirical. The debate has direct implications for whether the interpretive gap is closable by further scaling alone.

The second concerns heuristics and biases. Kahneman (2011) treats departures from normative rationality as systematic error; Gigerenzer and Gaissmaier (2011) reframe many such departures as ecological adaptations. The disagreement matters: if human interpretation is fundamentally heuristic and context-bound, models that assume normative reasoning will mispredict actual interpretive behaviour.

The third concerns interpretability itself. Lipton (2018) and Rudin (2019) argue that the field conflates transparency, post-hoc explanation, and human-grounded interpretability, with the result that systems may appear explainable without supporting any cognitive function for actual users. Miller (2019) sharpens the critique by showing that explanation research often proceeds without engaging the social science of how humans give and receive explanations. The contradiction is structural: explanation as a system property is conceptually distinct from explanation as a communicative act.

7. Toward Interpretation-Aware Systems

This section sketches a conceptual direction for future research. It is deliberately abstract. The argument of the paper is that interpretation should be treated as a research object in its own right; how that treatment is operationalised in particular systems is a separate matter that this paper does not adjudicate.

The conceptual move is to recognise that AI-mediated communication has at least two distinct layers of value. The first is the predictive or generative layer: the quality of the output a system produces. This layer is the focus of most contemporary evaluation, and significant progress has been made on it. The second is the interpretive layer: the perceived meaning that an output produces in a recipient given their context, prior

knowledge, and state. This layer is mediated by the cognitive and communicative phenomena reviewed in Section 5, and it is currently engaged with largely through proxies of the kind catalogued in Section 4, which the empirical literature shows to be dissociable from the cognitive states they are meant to represent.

An interpretation-aware paradigm, in the most general terms, would treat the second layer as a legitimate target of design, evaluation, and inquiry. The literatures reviewed in Sections 2 through 6 already provide substantial conceptual grounding for such a treatment, drawing on cognitive science, human–computer interaction, communication theory, and the trust and explanation literatures within AI itself. The shape of an interpretation-aware paradigm is a matter for further research; the case made here is for taking it up as a research direction.

What an interpretation-aware paradigm should not be is also worth stating. It should not be a relabelling of usability concerns, which would understate the conceptual move. It should not be a single architecture, since the receiver-side phenomena it engages with are heterogeneous and unlikely to admit a single computational treatment. And it should not be assumed to follow automatically from improvements in predictive or generative quality, since the empirical literature shows that these are dissociable.

8. Ethical and Research Implications

Several gaps and implications follow. First, established psychological frameworks (Cognitive Load Theory, framing, the elaboration likelihood model) are inconsistently integrated into the design and evaluation of AI-generated content. Their findings are stable, replicated, and decades old, but they are rarely operationalised in mainstream AI evaluation pipelines, which continue to rely on click-through, dwell time, and similar surface metrics as proxies for value. Second, the dissociation between behavioural engagement and underlying comprehension or trust is empirically documented (Avram et al., 2020; Lee & See, 2004) but is rarely treated as a problem to be measured rather than assumed away.

Third, the explainable-AI literature has identified the need for socially and psychologically grounded explanations (Miller, 2019) but few systems implement explanations validated against human interpretive performance in realistic contexts. Fourth, while simulation approaches (Park et al., 2023) show promise as tools for studying interpretation, their psychological validity is largely untested, and the conditions under which simulated populations approximate real ones remain unclear. Fifth, despite long-standing recognition

that meaning is context-bound and audience-dependent (Krippendorff, 2019; Hassenzahl, 2010), context-aware evaluation of AI outputs against specified audience profiles remains methodologically underdeveloped.

The ethical implications follow from the empirical observations. AI-mediated risk communication in domains such as healthcare, finance, and judicial decision-making shapes outcomes that matter at individual and societal scales; in such domains, interpretive failure is not a usability inconvenience but a determinant of welfare. The algorithmic-curation literature (Eslami et al., 2015; Avram et al., 2020) shows that interpretive failure at the level of information environments can shape civic and epistemic outcomes. Treating interpretation as a first-class concern is therefore not only a research agenda but an ethical commitment to taking seriously the conditions under which AI outputs become consequential for humans.

The implications for research direction are conceptual rather than prescriptive. Evaluation regimes that rely solely on accuracy, fluency, or engagement are likely to miss principal sources of failure in AI-mediated communication; they should be supplemented with research that engages the receiver side directly. The design of AI-generated outputs should explicitly account for the dual-process nature of human cognition. The move toward human-centred AI (Shneiderman, 2020; Shneiderman, 2022) implies that interpretive variability across audiences is a design parameter rather than an inconvenience; systems that assume a uniform recipient will systematically misjudge how their outputs land.

9. Conclusion

The peer-reviewed literature reviewed here supports the position that modern AI is highly capable at prediction and generation but insufficiently models human interpretation, defined narrowly as the receiver-side process through which a human constructs perceived meaning from a stimulus in context. The literatures synthesised do not articulate a single unanimous omission; they identify complementary gaps (causal, semantic, social, cognitive, communicative) that together motivate treating interpretation as a research object alongside the existing concerns of interpretability, understanding, explanation, comprehension, and trust.

Active research programmes (RLHF, user modelling, recommender systems, human–AI interaction, adaptive interfaces) already engage with the receiver side, and the paper's argument is that they do so through proxies that are useful but not equivalent to interpretation as defined here. The conceptual direction sketched

in Section 7 is offered as an invitation to a research agenda, not as a finished framework. Treating interpretation as a first-class concern does not require abandoning the predictive and generative capacities of contemporary AI; it requires recognising that those capacities are necessary but not sufficient, and that the conditions under which AI outputs become meaningful for human recipients warrant the same rigour of inquiry that the outputs themselves now receive.

References

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. https://doi.org/10.1145/3290605.3300233

Avram, M., Micallef, N., Patil, S., & Menczer, F. (2020). Exposure to social engagement metrics increases vulnerability to misinformation. *Harvard Kennedy School Misinformation Review, 1*(5). https://doi.org/10.37016/mr-2020-033

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-AI team performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 7*(1), 2–11. https://doi.org/10.1609/hcomp.v7i1.5285

Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., & Horvitz, E. (2019). Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(01), 2429–2437. https://doi.org/10.1609/aaai.v33i01.33012429

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 (FAccT '21)* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp.

5185–5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4299–4307). Curran Associates.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., & Sandvig, C. (2015). "I always assumed that I wasn't really that close to [her]": Reasoning about invisible algorithms in news feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 153–162). Association for Computing Machinery. https://doi.org/10.1145/2702123.2702556

Fischer, G. (2001). User modeling in human–computer interaction. *User Modeling and User-Adapted Interaction, 11*(1–2), 65–86. https://doi.org/10.1023/A:1011145532042

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62*, 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Hassenzahl, M. (2010). *Experience design: Technology for all the right reasons*. Morgan & Claypool.

Hassenzahl, M., & Tractinsky, N. (2006). User experience — a research agenda. *Behaviour & Information Technology, 25*(2), 91–97. https://doi.org/10.1080/01449290500330331

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kobsa, A. (2001). Generic user modeling systems. *User Modeling and User-Adapted Interaction, 11*(1–2), 49–63. https://doi.org/10.1023/A:1011187500863

Krippendorff, K. (2019). *Content analysis: An introduction to its methodology* (4th ed.). SAGE Publications.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences, 40*, e253. https://doi.org/10.1017/S0140525X16001837

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

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM, 61*(10), 36–43. https://doi.org/10.1145/3233231

Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. Pantheon Books.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences, 120*(13), e2215907120. https://doi.org/10.1073/pnas.2215907120

Norman, D. A. (2013). *The design of everyday things* (Revised and expanded ed.). Basic Books.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 27730–27744). Curran Associates.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)* (pp. 1–22). Association for Computing Machinery. https://doi.org/10.1145/3586183.3606763

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM, 62*(3), 54–60. https://doi.org/10.1145/3241036

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 123–205). Academic Press. https://doi.org/10.1016/S0065-2601(08)60214-2

Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender systems: Introduction and challenges. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook* (2nd ed., pp. 1–34). Springer. https://doi.org/10.1007/978-1-4899-7637-6_1

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE, 109*(5), 612–634. https://doi.org/10.1109/JPROC.2021.3058954

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction, 36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*(4), 295–312. https://doi.org/10.1016/0959-4752(94)90003-5

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453–458. https://doi.org/10.1126/science.7455683