

Before Judgment

How artificial intelligence can change the problem before a human

VERSION 1.0 / 30 AUGUST 2026

A human can retain the final decision while losing part of the independence that makes the decision meaningfully human. The loss may occur earlier, when artificial intelligence helps determine what is noticed, what is treated as relevant and what the situation is taken to mean.

Luiza Scurtu

Licence: Creative Commons Attribution-ShareAlike 4.0 International

DOI: [10.5281/zenodo.22177289](https://doi.org/10.5281/zenodo.22177289)

Abstract

Human oversight is commonly placed near the end of an artificial-intelligence-assisted decision: a system produces an analysis and a person reviews it before action. This assumes that the human reaches judgment with an independently formed account. Yet artificial intelligence can participate earlier by selecting information, organising relevance, proposing causes and shaping the language through which a problem becomes intelligible. Final authority may remain human while the decision object has been partly constructed elsewhere.

This article proposes the concept of judgment-object divergence. It develops one specific mechanism, latent-intent hypothesis divergence, in which humans and artificial-intelligence systems generate, retain and weight different explanations involving concealed goals, incentives, status, deception, fear, loyalty, retaliation or strategic conduct. Recent research already establishes the broader territory of epistemic mediation, changed problem formulation, representational gaps and weakened epistemic agency. The contribution here is to connect those findings to the behavioural hypotheses that shape ambiguous strategic situations and to show why final human approval may preserve authority in form while failing to preserve independent judgment in function.

A note on scope

The broad phenomenon—that artificial intelligence can reshape the environment in which people perceive, interpret and judge—is well supported. Judgment-object divergence and latent-intent hypothesis divergence are used here as organising concepts for a narrower mechanism.

Final human authority and upstream formation

Conventional accounts of responsible artificial intelligence place the human at the end of the sequence. A system produces an assessment; a person approves, rejects or modifies it. Because formal authority remains human, the process is described as human-led.

Before the recommendation appears, prior operations have established what counts as the problem, which signals and variables matter, which causal explanations are plausible and which uncertainties may be discounted. When artificial intelligence participates in those operations, final approval says little about independent human judgment.

The reviewer may understand and challenge the recommendation and retain the right to reject it. Yet a person can disagree within an artificial-intelligence-shaped account without recognising that a materially different account was possible.

The central issue is not confined to whether artificial intelligence makes the decision. It also concerns whether artificial intelligence has materially shaped what the decision is understood to concern.

A five-stage account of decision formation

A decision is better understood as a sequence than as a single moment. The stages overlap in practice, but separating them exposes where influence enters.

- **Situation or problem.** Something in the world requires attention, interpretation or action.
- **Perception.** Signals enter the working representation; other signals remain absent, invisible or discounted.
- **Interpretation.** The selected signals are organised into an account of what is happening, why it is happening and what matters.
- **Judgment.** The interpreted situation is evaluated against evidence, goals, values, risk and experience.
- **Decision and action.** A course is selected, authorised or enacted.

Governance usually concentrates on stages four and five. Yet the decisive divergence may occur at stages two and three, where perception and interpretation produce the operational reality that judgment receives.

A decision can be rational relative to its representation and still fail in the world. Reasoning may be coherent and evidence accurately cited; if the representation omitted a material signal, suppressed a plausible explanation or privileged the wrong account, good judgment is applied to the wrong object.

The observation behind judgment-object divergence

The originating observation arose during an analysis of a geopolitical agreement involving Venezuela. The human and artificial-intelligence participants had broadly similar public information. Their disagreement did not begin with different values or risk tolerances. It began with what each representation allowed to remain plausible.

The human account kept several unstated behavioural possibilities active: concealed self-interest, an official rationale used as cover for another objective, opportunism, power-seeking, historical conduct and the possibility that an actor would never state the operative intention publicly. The artificial-intelligence account initially assigned less weight to several of those possibilities because direct evidence of intention was incomplete.

This did not establish the human interpretation as correct. Concealed intent cannot be inferred merely because it is conceivable, and human judgment is itself vulnerable to over-attribution, stereotyping and unwarranted suspicion. The observation was narrower. By the time judgment began, the participants were applying judgment to materially different representations. One contained live hypotheses that the other had largely suppressed.

Judgment-object divergence occurs when people or systems appear to evaluate the same external situation but apply judgment to materially different operational representations because different signals, causal explanations, relevance weights or behavioural hypotheses entered earlier.

The decision object is not the external situation itself. It is the working representation of that situation: what the problem is believed to be, which actors and variables matter, what caused the present condition, which explanations remain credible and which uncertainties survive long enough to influence action.

Judgment-object divergence can occur between two humans, between departments, between institutions or within the same person over time. Artificial intelligence makes the condition especially important because one system can now influence problem formation at speed and scale while leaving final approval visibly human.

What judgment-object divergence is not

Judgment-object divergence is not simply automation bias or disagreement with an artificial-intelligence recommendation. Automation bias concerns excessive reliance on an output. Judgment-object divergence can exist even when the human critically evaluates or ultimately rejects that output, because the earlier representation of the problem may already have been shaped by the system. Nor is it simply information asymmetry: two participants can have access to the same facts while retaining different causal explanations, relevance weights and latent-intent hypotheses.

Latent-intent hypothesis divergence

Consequential situations are not always fully explained by documented facts and stated intentions. Actors manage impressions, protect status, conceal incentives, perform compliance, avoid retaliation or pursue private goals. Official language is evidence; it is not necessarily the complete causal account.

A latent-intent hypothesis is a revisable explanation involving an actor's unobserved goal, motive or strategic intention. It is neither proof nor accusation. It is a possibility retained because, if true, it would materially change how the situation should be interpreted.

Latent-intent hypothesis divergence arises when humans and artificial-intelligence systems differ in one or more of five operations: which motive hypotheses they generate; which they retain despite incomplete evidence; what probability or relevance they assign; what evidence causes them to update; and what would make them abandon the favoured explanation.

Latent-intent hypothesis divergence is therefore one possible source of judgment-object divergence, not its definition.

The difference may be subtle. One account treats a stated rationale as the operative explanation; another treats it as one piece of evidence. One excludes concealed intent because it cannot be directly observed; another keeps it alive but uncertain. A fluent narrative may compress ambiguity that another account preserves.

No factual error is required. The same facts can produce different decision objects when inclusion, relevance and explanatory weight differ. Fact-checking remains essential, but cannot resolve that difference.

Human fallibility and machine variability

The argument does not support a general claim of human superiority. Human beings overattribute disposition, rely on stereotypes, become anchored to first impressions and detect deception only modestly better than chance (Bond and DePaulo, 2006; Gilbert and Malone, 1995). Behavioural sensitivity can preserve a strategically important possibility, but it can also produce false attribution and unjustified confidence.

Artificial-intelligence performance is equally uneven. Large language models struggle on some tasks involving private information, white lies, complex social goals and motive generalisation (Mou et al., 2025; Liu et al., 2025; Shrestha et al., 2026; Zhou et al., 2025). Other work finds performance at or above human levels on bounded theory-of-mind and latent-motive tasks (Strachan et al., 2024; Fitch and Kvam, 2026).

A preprint posted on 26 August 2026 illustrates this variability. Sohail and colleagues compared humans and several large language models in strategic games, inferring trial-level beliefs about opponent sophistication. Performance differed sharply by model and prompting. In one task, GPT-5 tracked opponent sophistication more accurately than humans, while DeepSeek showed a weaker pattern (Sohail et al., 2026). The comparison therefore depends on the model, task, interaction design and form of inference.

The useful question is comparative: which hypotheses does each participant retain, how are they weighted and calibrated, and what errors follow? Preserving independent human formation prevents one fallible representational process from silently becoming the only process.

Position within the existing literature

Several literatures now converge on the proposition that artificial intelligence does more than provide answers inside a fixed epistemic environment. It can help construct the environment in which questions, evidence and judgment become possible.

Artificial intelligence as epistemic mediation

Pizarro Contreras, Zhao and Zhang describe conversational artificial intelligence as epistemic mediation that reorganises how people perceive, interpret and act. Their account of algorithmic capture concerns the production of what becomes visible and relevant before judgment takes place (Pizarro Contreras et al., 2026). Facchini's concept of mediational opacity goes further: artificial intelligence can reshape the epistemic environment while users remain unable to inspect how that reshaping occurred (Facchini, 2026).

Coeckelbergh locates the issue in epistemic agency, arguing that artificial intelligence can influence belief formation and revision in ways that diminish the person's capacity to act as an epistemic agent (Coeckelbergh, 2025). Kawall's account of epistemic autonomy supplies a useful foundation: autonomy includes meaningful control over topics of inquiry, sources, methods, goals and belief-formation processes, not merely the right to announce a conclusion after those choices have been made elsewhere (Kawall, 2024).

Problem formulation and downstream judgment

Research on strategic problem formulation provides direct evidence that timing matters. Wu, Kim and Lin found that introducing generative artificial intelligence during problem formulation changed the alternatives considered and reduced strategic focus compared with introducing it only during ideation. Their explanation points to artificial-intelligence-generated anchors entering the mental construction of the problem (Wu et al., 2025).

Kanis, Mann and Stumpf-Wollersheim similarly found that large language model use altered mental representations in a strategic task without reliably improving strategic foresight (Kanis et al., 2026). Cristofaro, Giardino and Augier show how generative artificial intelligence can reshape the construction, comparison and evaluation of strategic alternatives (Cristofaro et al., 2026). Together, these findings establish one component of the present argument: artificial intelligence can change how a problem is represented before a solution is selected.

Representational asymmetry

Yokoi and colleagues describe cognitive asymmetry in human–artificial intelligence collaboration. In their study of perfume creation, humans relied on tacit and embodied cognition while artificial intelligence operated through codified patterns. Effective collaboration required representational integration because the participants did not begin from equivalent representational frames (Yokoi et al., 2026).

Difference can create value. Artificial intelligence may generate a wider set of analogies while humans assess causal fit and contextual meaning (Sen et al., 2026). Complementarity depends on preserving distinct contributions. If artificial intelligence supplies the representation, interpretation and recommendation, the human may become an approver of coherence rather than an independent source of judgment.

Epistemic responsibility and upstream delegation

Theoharakis and Mylonopoulos distinguish augmentation, in which people retain evaluative responsibility, from substitution, in which reasoning is effectively delegated and outputs are accepted without sufficient independent evaluation (Theoharakis and Mylonopoulos, 2026). Ezra and Mishali distinguish the metabolisation of artificial-intelligence-generated material into genuinely owned thought from epistemic seduction, where fluency allows the user to bypass that work (Ezra and Mishali, 2026).

The ethics of artificial-intelligence scribes provides a particularly concrete analogue. Ursin and Salloch argue that when a system controls how a clinical encounter becomes a record, it shapes the evidential object available to the physician. The physician risks becoming an editor of an artificial-intelligence-constructed representation while retaining formal responsibility for the clinical decision (Ursin and Salloch, 2026).

In strategic reasoning, van Rooyen and Nandhakwang's preprint identifies machine-mediated intent attribution, the leap from observable footprint to motive and interpretive dependency on machine-generated accounts (2026). It comes closest to the latent-intent problem considered here. The remaining issue is how such attribution changes the object later presented to human judgment.

The limits of final-stage human oversight

Traditional oversight asks whether the human can understand, challenge or reject the artificial-intelligence output. Those are necessary questions. They are not sufficient when the output is the end of a representational process that began earlier.

An explanation can show why a system reached a conclusion from the variables present in its representation. It cannot reveal every variable that never entered, motive hypothesis that was never generated or framing suppressed before the explanation began. Challenging an absent possibility is difficult once a fluent account has established the vocabulary and boundaries of the problem.

Explainability operates downstream of representation formation; judgment-object divergence directs attention to the formation of the representation itself.

Research on human oversight has already shown that formal review can become rubber-stamping when people lack time, epistemic access, causal power or a reason to challenge the system (Green, 2022; Sterz et al., 2024; Zhu et al., 2026). Cognitive forcing can reduce overreliance; requiring an initial human judgment before displaying advice is one effective example, although people often dislike the friction it introduces (Buçinca et al., 2021).

Judgment-object divergence moves the safeguard further upstream. The relevant question is not only whether the person independently evaluated the recommendation. It is whether the person independently formed, or was able to reconstruct, the reality upon which the recommendation operates.

Formal authority answers who is allowed to decide. Functional independence asks who helped construct the reality that the decision treats as true.

Principles for functionally independent judgment

Meaningful human judgment does not require excluding artificial intelligence from problem formulation. Artificial intelligence can extend alternative generation, pattern detection and analytical reach. The process must prevent assistance from becoming an unexamined source of representational authority. Six principles follow.

- **An independent first account.** Where stakes and ambiguity justify it, the human should record an initial formulation before seeing an artificial-intelligence-generated narrative: what happened, what matters, what remains uncertain and which explanations are still live.
- **Decision-object provenance.** The process should show how the working representation was constructed: which sources entered, which were excluded, who assigned relevance and where artificial intelligence added interpretation rather than evidence.
- **Rival representations.** The system should generate more than one coherent account, including non-motive explanations and explanations that would contradict the preferred interpretation.
- **Separation of evidence and intention.** Observable facts, causal inferences, stated rationales and latent-intent hypotheses should remain distinct. A plausible motive should never quietly change grammatical form and reappear as a fact.
- **Preserved uncertainty.** The process should retain what is unknown, what would disconfirm the favoured account and what new evidence would justify an update.
- **Reconstruction before approval.** The reviewer should be able to rebuild the problem from sources and alternatives, not merely edit the final narrative or choose among options already framed by the system.

These principles introduce friction. In consequential decisions, such friction can make independence observable. A process optimised exclusively for speed and fluency tends to compress competing representations into the most usable account, although usability and fidelity are not equivalent.

Boundary conditions and domains of application

Judgment-object divergence does not explain every disagreement with artificial intelligence. It matters most where evidence is incomplete, actors are strategic, motives are weakly observable, causal accounts are contested and interpretation can materially alter the available action.

That includes geopolitical analysis, organisational politics, negotiations, investigations, restructuring, acquisitions, regulatory strategy, leadership assessment and other environments where stated positions coexist with incentives and concealed constraints. It may also matter in clinical, legal or

safeguarding settings when documentation and interpretation determine what later reviewers are able to see.

The concept has less explanatory value in tightly bounded tasks with stable categories, complete information and clear ground truth. It should remain confined to settings in which behavioural interpretation can materially change the operational representation.

A pre-judgment diagnostic

Before relying on a consequential artificial-intelligence-assisted judgment, the following questions expose whether the decision object has become shared, displaced or opaque:

- What entered the frame, and what remained outside it?
- Which parts of the account are observations, which are inferences and which are claims about intention?
- Who decided what was relevant?
- Which explanations were never generated or were dismissed early?
- Did an official rationale become the causal account without comparison against incentives and conduct?
- What uncertainty was preserved, and what was converted into a confident narrative?
- Did the human form an account before exposure to the artificial-intelligence framing?
- Could the human reconstruct the problem from sources and rival explanations without relying on the system's narrative?
- Would a materially different representation lead to a different judgment or action?

The purpose is to make the formation of the decision object sufficiently visible to challenge. A disciplined process must be capable of introducing a concealed-intent hypothesis when the evidence warrants it and removing that hypothesis when the evidence does not sustain it.

Governance implications

The phrase human in the loop places the safeguard too late. A loop can contain a person while delivering an artificial-intelligence-constructed reality for approval. Presence is not independence, and accountability does not show who formed the account.

Governance therefore needs to distinguish three questions. Who has formal authority? Who constructed the operational representation? Who bears the consequence if that representation is wrong? These answers are already separating in many organisations; artificial intelligence accelerates and obscures the separation while a final human signature legitimises it.

A stronger standard asks whether the human contribution remained epistemically independent when the problem was formed. It requires an independent source of framing, rival explanation and disconfirmation before the organisation treats the resulting object as decision-ready.

Conclusion

Artificial intelligence increasingly selects, summarises, organises, interprets and explains before the formal decision begins. Those functions help determine the reality to which judgment will be applied.

Existing research establishes several forms of upstream artificial-intelligence influence. Judgment-object divergence provides a conceptual account of what happens when those processes cause people and systems to arrive at judgment with materially different operational representations, including through differently weighted latent-intent hypotheses.

A proportionate response keeps representation formation visible, rival accounts available, evidence separate from motive attribution and the human able to reconstruct the problem rather than merely approve its final rendering.

Final authority may remain human even when artificial intelligence has already materially shaped the object over which that authority is exercised.

References

- Bond, C. F., Jr., and DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, 10(3), 214–234. [Source](#)
- Buçinca, Z., Malaya, M. B., and Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on artificial intelligence in assisted decision-making. *Proceedings of the Association for Computing Machinery on Human-Computer Interaction*, 5, Article 188. [Source](#)
- Coeckelbergh, M. (2026; online 2025). AI and epistemic agency: How AI influences belief revision and its normative implications. *Social Epistemology*, 40(1), 59–71. [Source](#)
- Cristofaro, M., Giardino, P. L., and Augier, M.-S. (2026). Generative AI and cognitive biases in strategic decision-making. *Strategic Organization*, online first. [Source](#)
- Ezra, R., and Mishali, M. (2026). The relational–epistemic stance: Generative AI as a dynamic transitional object. *AI & Society*, 41, 5027–5042. [Source](#)
- Facchini, A. (2026). Mediatonal opacity: How AI reshapes our epistemic environments, and how to audit it. Working paper; accepted at the 2026 Conference on Artificial Intelligence, Ethics, and Society. [Source](#)
- Fitch, A. K., and Kvam, P. D. (2026). Cognitive models facilitate real-time inference of latent motives. *Scientific Reports*, 16, Article 6444. [Source](#)
- Gilbert, D. T., and Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, 117(1), 21–38. [Source](#)
- Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 45, 105681. [Source](#)
- Kanis, T., Mann, J. E., and Stumpf-Wollersheim, J. (2026). AI-augmented strategic decision-making under time constraints: An experimental study on mental representations and strategic foresight. *Strategy Science*, 11(1), 75–92. [Source](#)
- Kawall, J. (2024). Epistemic autonomy and the shaping of our epistemic lives. *Social Epistemology*, 38(3), 374–391. [Source](#)
- Liu, Y., Pretty, E. J., Huang, J., and Sugawara, S. (2025). TactfulToM: Do large language models have the theory-of-mind ability to understand white lies? *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 25043–25061. [Source](#)
- Mou, X., et al. (2025). AgentSense: Benchmarking social intelligence of language agents through interactive scenarios. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4975–5001. [Source](#)
- Pizarro Contreras, R., Zhao, Z., and Zhang, G. (2026). Just in time: The algorithmic capture of truth and opportunity in AI-reconfigured judgment. *Humanities and Social Sciences Communications*, 13, Article 1074. [Source](#)
- Sen, P., Workiewicz, M., and Puranam, P. (2026). Can LLMs aid analogical reasoning for strategic decisions? A comparative study. *Strategy Science*, 11(1), 118–136. [Source](#)
- Shrestha, K., Gupta, A., Dukkipati, H., Shi, Z., and Mataric, M. (2026). Can large language models infer human actions and motives? Evaluation in social prediction and inspection games. *Findings of the Association for Computational Linguistics: ACL 2026*, 32791–32803. [Source](#)
- Sohail, A., Zhong, X., Konovalov, A., Lockwood, P. L., and Zhang, L. (2026). Assessing mentalization in humans and large language models. Preprint. [Source](#)

- Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P., and Langer, M. (2024). On the quest for effectiveness in human oversight: Interdisciplinary perspectives. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 2495–2507. [Source](#)
- Strachan, J. W. A., et al. (2024). Testing theory of mind in large language models and humans. Nature Human Behaviour, 8, 1285–1295. [Source](#)
- Theoharakis, V., and Mylonopoulos, N. (2026). AI epistemic disengagement and consumer dependence: An augmentation–substitution framework. Psychology & Marketing, 43, 1675–1686. [Source](#)
- Ursin, F., and Salloch, S. (2026). The ethics of AI scribes as epistemic agents. JMIR Medical Informatics, 14, e88235. [Source](#)
- Vaccaro, M., Almaatouq, A., and Malone, T. W. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. Nature Human Behaviour, 8, 2293–2303. [Source](#)
- van Rooyen, J., and Nandhakwang, N. (2026). From footprint to motive. SSRN working paper. [Source](#)
- Wu, N., Kim, H., and Lin, C. (2025). From problems to solutions in strategic decision-making: The effects of generative artificial intelligence on problem formulation. INSEAD Working Paper 2025/42/STR. [Source](#)
- Yokoi, T., Laureiro-Martinez, D., Magni, F., and Brusoni, S. (2026). Organizing across cognitive asymmetry in human–AI collaboration: A study of perfume creation. Strategic Management Journal, 1–25. [Source](#)
- Zhou, J., et al. (2025). SocialEval: Evaluating social intelligence of large language models. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, 30958–31012. [Source](#)
- Zhu, L., Lu, Q., Ding, M., et al. (2026). Designing meaningful human oversight in AI. AI and Ethics, 6, Article 286. [Source](#)