Should We, Like Artificial Intelligence, Sometimes Rewrite Our Thinking From Scratch?

What Separates Machine From Human

Ramin Saadat
Independent Researcher; Founder, British Institute for Ignorance Studies (BIIS) (biis.uk)

AI Disclosure and Translation Statement:
The original manuscript of this work was authored in Farsi (Persian) by the researcher. The English version was produced through AI-assisted translation and linguistic refinement, utilizing Google Translate and Large Language Models. The author has meticulously reviewed, edited, and verified the final English text to ensure that all theoretical nuances, philosophical arguments, and conceptual frameworks accurately reflect the original Persian content. The author remains fully responsible for the integrity and accuracy of the translation. The original Persian manuscript is available from the author upon reasonable request.

Abstract

This article begins with a simple observation: midway through a long conversation, an artificial intelligence sometimes suggests starting the chat afresh. Rather than taking the model's own explanation for this suggestion at face value, the article turns to an external measure of performance. Drawing on Laban et al. (2025), it shows that when information reaches a language model piecemeal rather than all at once, the resulting drop in performance comes mainly not from reduced capability but from instability: the model anchors on whatever arrives first. The article then asks whether we are entitled to extend this finding to human thinking, and, drawing on an independent literature on anchoring and adjustment in cognitive psychology (from Tversky and Kahneman to recent work on language models), shows that such licence exists, at least at the level of behavioural pattern. In humans, however, this anchor can become attached to something the machine lacks: a self. From here, two layers are distinguished: a cold layer, shared by model and human, and a warm layer, found only in humans, in which the initial anchor turns into identity-linked commitment. The article's practical conclusion is that neither the inner sense of steadfastness nor the act of questioning oneself offers a reliable test for distinguishing sound conviction from premature commitment. Only a behavioural test, independently reconstructing the problem and comparing the result, can make that distinction.

Keywords: premature commitment; anchoring and adjustment; narrative commitment; turn-based information structure; multi-turn conversations with artificial intelligence; belief perseverance; cold and hot cognition

---

Introduction

Many of us have experienced this scene: midway through a long conversation with an artificial intelligence, we ask whether it would be better to start the chat afresh. The model answers yes, and in a single line explains why. It says that keeping track of something said piecemeal, across the course of the conversation, has become harder for it, and that it might be better to reformulate everything said so far in one go, from scratch.

This explanation appears simple, but it is worth pausing over. The model does not say ‘I made a mistake’ or ‘I no longer know.’ It says something more precise: the problem lies not in the content

of the conversation but in how the information arrived, piecemeal and over time rather than all at once. This is a claim about structure, not about knowledge itself, and it is precisely this feature that makes it testable: one can give the model the same content once piecemeal and once all at once, and see whether the result differs. This precision forms the core of the article: if the order in which information arrives, independent of the information itself, affects the outcome, does something similar happen in our own thinking? And does the model's simple answer hold a lesson for how we think and reason?

This article calls this feature turn-based information structure (turn-based structure, sharded versus concatenated): whether a request, or a body of evidence, reaches the model complete and in one piece, or piecemeal, across successive turns of a conversation. The important point is that this is a feature of how the content is arranged, not of the content itself. If the same request, carrying the same volume of information, is given to the model once in full and once piecemeal, the model's final answer can differ. Throughout the article, the term retains this specific meaning.

At this point, however, one point needs clarifying: we cannot accept this brief answer without question. The answer is, after all, the model's own report of its internal state, and such a report is not reliable evidence for any firm conclusion about its actual internal workings. It is more likely a repetition of a pattern learned from training data. An internal report, whether from a model or from a human, is weak evidence for understanding and explaining the real internal mechanism.

The argument of this article about the resemblance between machine and human, and about what we might learn from the AI's response, cannot therefore rest on the model's verbal answer. It must rest instead on an external measure of the model's performance, not on what it tells us.

With this caution in mind, we must ask: is what the model points to merely a technical limitation, or a door onto a question human thought has long wrestled with under other names, posed in various forms, and now reopened in this one? How far does our adherence to a view signal that it is correct, and at what point does it signal something else?

This adherence and persistence, which this article calls narrative commitment, is more than a structural fact. It is a sense of responsibility towards one's own past statements and an identity continuity that links those statements to a single self over time. This commitment has two faces. Its healthy face is exactly this responsibility and continuity; its other face is locking onto an initial position and rationalising it. This article is concerned only with the second face.

By the end, this article arrives at a test, but a behavioural one, not an inner feeling. The feeling in question is the one that arises after a period of thinking, when we no longer weigh new evidence. This feeling alone is not sufficient evidence; we must also run the test itself: write the problem out afresh, or explain it to someone with no prior knowledge of it, and see whether we reach the same conclusion again. If the result differs, we can conclude that what we took for fidelity to the truth has probably given way to premature commitment to our own assumptions.

To pursue these questions and tests, the article proceeds in four sections. The first looks at what happens in experiments with language models. The second asks whether we are even entitled to extend this finding to humans, and examines evidence bearing on that question. The third turns to the part of the picture that belongs to humans alone. The fourth asks what can be done, if this diagnosis is correct.

Let us now look at the experiment itself: what exactly happens when a language model loses its way in the middle of a conversation?

---

1. What Happens in the Model

Laban and colleagues, from Microsoft Research and Salesforce Research (Laban et al., 2025), answered exactly this question. Their question was whether the final answer would differ if the same instruction, with the same content and the same volume, were given to a model once in full and once piecemeal across a conversation.

To test this, they gave the model a single instruction in two forms: once complete and in one piece, and once broken into several shards, each turn revealing only one of them. In the second form, the model had to decide, before the problem was complete, what to say: ask a question, guess, or give a final answer.

The experiment involved more than two hundred thousand simulated conversations, six task types (coding, database querying, API calls, mathematics, data-to-text description, and summarisation), and fifteen models. Average performance in the sharded condition was thirty-nine per cent lower than in the full condition. This drop appeared across all the models, and showed up even in two-turn conversations (a conversation with only two turns: one user message, one model response; the technical literature calls each message or response a turn, not each exchange).

The more important finding, however, emerged once this drop was decomposed. Each request was run ten times, and two separate measures were taken: aptitude, the ninetieth-percentile score across those ten runs, and unreliability, the gap between the ninetieth and tenth percentile of the same runs. Aptitude fell by only sixteen per cent: at its best, the model retained almost the same power it had shown in the full condition. Unreliability, however, rose by one hundred and twelve per cent. In other words, the problem was that the model had become unreliable, not that it could no longer reach the level of performance it showed in the full condition.

Why did this happen? The researchers identified four separate behavioural patterns in the model's own responses, each with its own evidence. First, the model reaches a final answer early and fills in unstated details with assumptions; the accuracy of these early answers was only 30.9 per cent, against 64.4 per cent when the model waited until the information was complete.

Second, the model relies excessively on that early answer, even when it is wrong.

Third, it weights the first and last turns more heavily than the turns in between; references to middle turns fell below twenty per cent.

Fourth, its answers grow longer as the conversation progresses, by twenty to three hundred per cent, and this verbosity crowds out the user's own words.

The researchers summed up their finding this way: once the model goes down the wrong path in a conversation, it gets lost and does not find its way back.

This article calls the first and second causes premature commitment: one is reaching an answer early, which Laban calls premature answer attempts; the other is excessive reliance on that answer, a consequence of the first, which Laban calls over-reliance on prior attempts. The term premature commitment itself does not come from Laban's article; it is ours.

This finding has its limits. The user in the experiment was itself a model, not a human; the tasks were all analytical and in English; and in a supplementary task tested separately, outside the six main tasks, which could be done piecemeal, namely sentence-by-sentence translation, no drop appeared. The authors speculate that the drop occurs in generative tasks that are complex and non-decomposable.

The conclusion to be drawn from this research is thus limited but solid: in a system where turn-based information structure can be altered independently of aptitude, most of the drop in performance comes from unreliability, that is, from the spread of outcomes across different runs of a single request, rather than from reduced aptitude, meaning a fall in the ceiling of the model's best possible performance.

This finding also offers a bridge to the human case that we will need later. If the drop comes mainly from unreliability rather than one-sided deviation, this means that premature commitment does not drive every run of a request towards one fixed outcome. Each run anchors on whatever shard arrived first, and because that shard differs from run to run, the outcomes scatter rather than deviate uniformly. In other words, the behavioural signature of anchoring, across a population of runs, is dispersion, not a bias towards one fixed direction.

With this picture of the model in hand, the next natural question is whether we are entitled to extend a finding obtained about a computational system to human thinking at all.

---

2. The Cold Layer: Are We Entitled to Compare?

Given this picture of the model drawn in the previous section, the next natural question is whether we are entitled to extend a finding obtained from studying a computational system to human thinking at all, given that machine and human are not alike in mechanism. The machine's knowledge and reasoning capacity, once trained, remain fixed and do not change over the course of any conversation; it has neither a memory of past conversations nor an identity to preserve. What changes with each turn is only the input text in front of it. The human, by contrast, carries a memory and an identity through time and attributes past statements to that identity. This is an important and real difference, and it cannot and should not be concealed.

In no field of science, however, does the right to compare depend on two systems sharing the same mechanism. When ethologists compare problem-solving in crows and in humans, for instance, the crow brain and the human brain share no structural resemblance at all; the researcher's licence to compare comes from elsewhere, from an independently documented behavioural pattern found in both. If such a pattern, documented in two entirely separate literatures, one on language models and one on human cognition, has been recorded before this article and independently of it, that alone licenses comparison at the level of behaviour, not a claim of shared mechanism. Let us see whether such a pattern exists.

Borrowing from the cold-cognition and hot-cognition distinction in psychology, this article calls this layer the cold layer. Cold cognition is processing that occurs free of emotion and personal interest, purely informational and analytical. Hot cognition is processing in which emotion, motivation, or the interests of the self are involved.

Here too, then, the cold layer is processing that needs no self whose reputation or identity is at stake; it is simply asymmetric weighting of the order in which information is received.

One relevant explanatory framework in cognitive psychology is anchoring and adjustment, demonstrated by Tversky and Kahneman in 1974 (Tversky and Kahneman, 1974). On this account, judgement starts from a point, the anchor, and moves towards a more correct answer through adjustment. This adjustment typically stops too early, however, and the final answer still bears the trace of that starting point.

In the view of this article, part of what was observed in conversation with the model can also be read within this Tversky and Kahneman framework: whatever arrives first becomes the anchor for the rest of the way. Laban and colleagues did not draw this connection themselves; what they reported was simply the model's behaviour, and this reading, which connects that behaviour to anchoring theory, is this article's own interpretation.

Dennis and Ahn tested this effect in causal relationship learning in 2001 (Dennis and Ahn, 2001). In their study, participants saw identical evidence in two different orders, and their final judgements differed. A hypothesis is formed from the earliest evidence, and later evidence corrects it, but by less than the content of that evidence alone would warrant. In an ideal Bayesian update, if the evidence is conditionally independent and the probability model remains fixed, the order in which data arrives should not matter; Hogarth and Einhorn (1992) formalised this deviation as the belief-adjustment model.

This boundary calls for caution, however. Marsh and Ahn (2006) showed that the primacy effect does not always occur. It depends on whether the mind has the chance to form a single hypothesis from the outset and stick with it. When that chance is removed, for instance because a person must distinguish several causal relationships at once, the mind can no longer lock easily onto an initial hypothesis, and the primacy effect weakens; this capacity is also linked to a person's verbal memory. More interestingly, under the same conditions, a recency effect sometimes appears instead of a primacy effect: more weight is given to the most recent evidence rather than the earliest. This shows that sensitivity to order is a stable phenomenon, but its direction is not fixed.

In other research, this same family of effects has been tested in language models themselves. Wang and colleagues showed in 2023 that in classification tasks, ChatGPT favours labels that appear earlier in a list, even when the content is held constant and only the order changes (Wang et al., 2023).

Then, in 2024, Guo and Vosoughi examined serial position effects across more models and wrote that comparing model behaviour with human cognition suggests this sensitivity may be more than an engineering issue, and may relate to the nature of the attention mechanism itself (Guo and Vosoughi, 2024).

Next, in 2025, Hämäläinen replicated Asch's classic experiment (Asch, 1946) on three commercial models, with mixed results (Hämäläinen, 2025). In one test, ChatGPT favoured the candidate whose positive traits were listed first; in another, whenever the models showed a preference, they leaned towards the candidate whose negative traits were listed first. This result is evidence of sensitivity to order, but not evidence of a human-like primacy effect.

Laban's own research offers a further clue, though an ambiguous one: in that experiment, models weighted the first and last turns more heavily than the middle turns, showing both primacy and recency together, not primacy alone. This pattern is therefore not pure evidence of anchoring, but it does at least show that sensitivity to order goes beyond a one-sided deviation. Some researchers (Guo and Vosoughi, 2024) suspect that part of this sensitivity to order in models is a product of how they were trained, rather than an inherent feature common to all of them.

It must also be said that these studies concerned only the order of complete information, whereas in Laban's research, information arrives incomplete and gradually. The two phenomena are related, but not identical.

We can now return to the question that opened this section: were we entitled to compare? Given everything reviewed here, the answer seems to be yes, at the level of behavioural pattern, since this same cold layer is documented in both model and human and requires no motive.

This, however, is only half the human picture. The other half is the warm layer, which the machine lacks, and the next section turns to it.

---

3. The Warm Layer: When the Anchor Becomes Attached to Identity

The previous section showed that the cold layer, anchoring on whatever arrives first, is documented in both model and human, and requires no motive or identity. In humans, however, something further completes the picture: whatever forms the hot side of the cold-and-hot-cognition distinction, processing that is no longer neutral and purely informational but involves emotion, interest, and, most importantly, the self. When the initial anchor becomes attached to this self, something more than a simple cognitive error arises: a commitment that drives the person to defend it, even when the evidence points the other way.

In 1976, Staw demonstrated this in a structured experiment. Two hundred and forty management students were placed in the role of a company director who had to divide a research and development budget between two company divisions. Half the participants had made this initial decision themselves; the other half were simply told that the decision had already been made. Everyone was then told that the division receiving the funding had performed poorly, and asked to allocate the budget again. Those who had made the initial decision themselves gave more money, this time too, to that same underperforming division, considerably more than those who had not made the decision themselves (Staw, 1976).

Staw explained this through people's need to justify their own decisions. Once responsibility for a choice becomes attached to the self, the failure of that choice is no longer merely a technical error; it is a failure of the self, and the way to defend the self becomes investing further in the same path.

Ross, Lepper and Hubbard tested this mechanism from another angle (Ross, Lepper and Hubbard, 1975). In their experiment, participants were tested on distinguishing genuine suicide notes from fake ones, but the feedback they received was random, not real; some were told their performance was excellent, others that it was poor. Everyone was then told that this feedback had been entirely fabricated and bore no relation to their actual performance. Even so, participants' beliefs about their own ability remained influenced by that fake feedback.

The researchers themselves attributed this to a feature of the experimental design, and later literature named the phenomenon this research gave rise to belief perseverance. Before hearing that the feedback was fake, participants had been asked to write down a reason for their (fabricated) performance (‘because I am good at maths’ or ‘because I am very careful’). Those who had written such reasons held onto their beliefs far more stubbornly after the feedback was discredited than those who had not gone through this step. It is worth remembering that this experimental evidence supports exactly this claim, and no more: what kept the belief alive was not the feedback itself, but the explanation the person had constructed for it.

Kruglanski (1989; Kruglanski and Webster, 1996) formulated this pattern within a broader theory, the need for closure. On his account, the mind faced with ambiguity does two things. First it seizes, clinging to the first acceptable explanation to escape the discomfort of indecision. Then it freezes, resisting revision of that explanation, because revision means returning to that same unwelcome

uncertainty. The greater a person's need for closure, for instance under time pressure or fatigue, the stronger this seizing and freezing becomes.

Luchins showed this same persistence in 1942, even with no emotional or social content at all, in one of the simplest and best-known experiments in cognitive psychology (Luchins, 1942). Participants were given three water jars, holding 21, 127 and 3 units, and asked to measure out 100 units. The solution was to fill the 127-unit jar, then empty it once into the 21-unit jar and twice into the 3-unit jar, leaving 100 units. After several similar problems solvable with this same formula, participants were given a problem that could still be solved this way, but also had a much simpler solution: simply adding two of the jars together. The group that had grown accustomed to the complex formula used the same long method again and never saw the simple one; a control group, without that earlier practice, found the simple solution immediately. Once a method has worked, the mind continues with that pattern, even when a simpler route lies in plain sight, an effect known as the Einstellung effect.

As we can see, none of these four studies concerns machines; their subject is exactly the motive and identity the model lacks. There is, nonetheless, a resemblance at the level of output. In Laban's research, the model relied excessively on its own earlier, incorrect answers. In Ross and colleagues' research, the human remained loyal to an explanation they had constructed themselves. In Staw's research, the human remained loyal to a decision they had made themselves.

The cause of this persistence is not the same in the two systems, however. In the model, it appears to be the asymmetric weighting of what came earlier in the text; in the human, it is the need to justify and defend the self. This, too, should be kept at the level of output resemblance, not taken as evidence of a shared cause.

Taken together, then, what we have seen in these two sections shows that we are dealing with two distinct layers. The cold layer, anchoring on whatever arrived first, is documented in both model and human and requires no motive. The warm layer, rationalisation and identity preservation, exists only in humans, and turns the initial anchor into commitment.

What the two share, however, is that in both cases, adherence to the initial path can take on the appearance of stability and correctness without actually being correct. If this diagnosis is right, the next question is practical: once the initial anchor has turned into commitment, what can be done?

The next section will show that Laban's answer for the model offers a clue, but does that same clue work for humans too?

---

4. What Can Be Done?

In answer to the question raised at the end of the previous section, Laban and colleagues propose a practical measure for models: gather everything said piecemeal across a conversation into a single complete request, and begin the conversation afresh with that.

Their own experiment supports this recommendation: they observed that the consolidated request restored, on average, around ninety-five per cent of full-condition performance.

They also tested two intermediate remedies that compensate within the same conversation, without restarting it: restating the entire history at the final turn, and repeating it incrementally at each turn.

The second of these, however, recovered only fifteen to twenty per cent of the drop in the two models tested, while the first gave better results, although in a real conversation one cannot know

in advance which turn will be the last. The authors themselves call these remedies mitigations, not principled solutions.

In determining the human equivalent of this measure, however, we must be precise.

Laban's remedy is consolidation, not deletion. No evidence is discarded; only the historical order in which it arrived is removed. The correct human equivalent is the same: we take all the evidence, arguments and sources we have gathered so far, irrespective of the order in which they reached us, and reformulate them in one go, then reason from that fresh formulation. This is not a matter of refusing to look at them.

Explaining the problem to someone with no knowledge of our back-and-forth thinking does this naturally, for example, since it forces us to restate everything we know in one orderly pass. It therefore comes closer to Laban's remedy.

Writing on a fresh sheet of paper can play the same role, provided the aim is a complete reconstruction of the evidence, not forgetting it. This does not erase memory, which is not possible for humans, but it does place a distance between the problem itself and the history of commitments we have made to it.

This technique is related to several well-known methods in decision science, although it is identical to none of them. It bears a resemblance, for instance, to the outside view, proposed by Kahneman and Lovallo in 1993 (Kahneman and Lovallo, 1993), which judges by comparison with similar past cases rather than by reconstructing the evidence itself.

It also resembles the premortem, proposed by Klein, which uncovers blind spots before they occur by imagining a hypothetical failure (Klein, 2007).

What all three share is the creation of distance between the current judgement and the framework within which the initial decision was formed.

What is genuinely new here is that in a language model, turn-based information structure can be manipulated directly and its effect on performance measured; in humans too, the effect of order and manner of presentation can be tested experimentally, as the research reviewed in the previous sections shows. Completely removing a person's cognitive history and prior commitments, however, so that no trace remains, is a different matter and has never been possible in humans.

It must also not be forgotten that Laban's experiment concerned the model. Whether its human equivalent has the same effect is a testable hypothesis and proposal, not a finding one can claim.

How, then, do we know when it is time to do this?

The first thing that comes to mind is the feeling described at the start of the article: a moment arrives when we no longer weigh new evidence and merely try to fit it to what we have already said. This feeling alone is not enough, however, not only for the general reason given earlier, but for a more specific one: asking ourselves ‘if I were starting from scratch today, would I reach the same conclusion?’ is itself an inner judgement carried out, and answered, by that same committed mind. Ross, Lepper and Hubbard's research showed exactly this in humans: when people were asked to imagine that the original basis of their belief had been fabricated, the belief remained in place. This question should therefore not be treated as a test, only as a prompt to act. The test must be behavioural and external: write the problem out on a fresh sheet of paper, or explain it to someone with no knowledge of our prior thinking, and see whether the fresh formulation and conclusion match what we said before. Unlike questioning ourselves, this produces an output that can be compared with the earlier text.

It must also be said, however, that not all adherence is premature commitment. Sometimes a position that survives this comparison is a sound conviction, reached through independent thought and sustained reflection, and this conviction needs no renewal. What should be renewed from time to time is not this sound conviction itself, but the habit of testing it. In other words, even a belief that currently seems correct is worth putting to the test again occasionally, from scratch, on a fresh sheet of paper, just as one must sometimes start a fresh chat with an artificial intelligence.

Conclusion

This article began with a familiar scene: a model that, midway through a long conversation, suggests starting the chat afresh. Our question was what lesson this behaviour holds about the actual workings of our own minds.

Directly measuring the machine's behaviour revealed something that had nothing to do with being a machine: the order in which information arrives, on its own and apart from the content of that information, changes the final outcome. Placed alongside an independent literature on human cognition, this finding clarified one thing: this vulnerability is not unique to the machine. We, too, give asymmetric weight to whatever arrives first, in the same way and perhaps through the same, as yet unknown, mechanism.

The more important point, however, was that for the machine, correcting this flaw is cheap. Nothing in the machine is harmed by the correction, because the machine has nothing it needs to defend. This is not so for the human. When that same initial anchor becomes attached to the self, correction is no longer merely a change of opinion; it becomes a threat to the very person who formed that opinion. For this reason, precisely where we most need to reconsider, we also show the most resistance to doing so. This, rather than the more superficial resemblance in anchoring, is what this article has really shown about the human, set against the machine.

From this difference follows a practical, and not entirely welcome, conclusion. If we cannot trust our own inner sense of a belief's correctness, not for the simple reason that we might be wrong, but for the more precise reason that once a belief is attached to identity, that sense is no longer a neutral report of a judgement but a product of the very thing it is meant to assess, then the only way forward is to bypass this sense. Not by questioning ourselves, but by producing something external and comparable. What we borrow from the machine is not what it says, but how it tests itself.

Yes, we too must sometimes rewrite our thinking from scratch, not because our beliefs, like long conversations, decay of their own accord, and not because we should follow the machine's example of restarting with ease. The reason is the opposite: because starting again is, unlike for the machine, difficult and costly for us, and this very difficulty is the only means of distinguishing sound conviction from premature commitment.

The machine can rewrite with ease, because it has nothing that is harmed by writing again. Rewriting is difficult for us, however, because whatever we have written has become part of who we are. This very difficulty of rewriting is a sign of its necessity, not an excuse to avoid it. True fidelity lies not in holding onto what we have already written, but in remaining always ready to write it again.

References

Asch, S. E. (1946). Forming impressions of personality. Journal of Abnormal and Social Psychology, 41(3), 258–290.

Dennis, M. J., & Ahn, W.-K. (2001). Primacy in causal strength judgments: The effect of initial evidence for generative versus inhibitory relationships. Memory & Cognition, 29, 152–164.

Guo, X., & Vosoughi, S. (2024). Serial position effects of large language models. arXiv:2406.15981.

Hämäläinen, M. (2025). On psychology of AI — Does primacy effect affect ChatGPT and other LLMs? arXiv:2504.20444.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. Cognitive Psychology, 24, 1–55.

Kahneman, D., & Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk taking. Management Science, 39(1), 17–31.

Klein, G. (2007). Performing a project premortem. Harvard Business Review, 85(9), 18–19.

Kruglanski, A. W. (1989). Lay epistemics and human knowledge: Cognitive and motivational bases. Plenum Press.

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: “Seizing” and “freezing”. Psychological Review, 103(2), 263–283.

Laban, P., Hayashi, H., Zhou, Y., & Neville, J. (2025). LLMs get lost in multi-turn conversation. arXiv:2505.06120. Published as a conference paper at ICLR 2026.

Luchins, A. S. (1942). Mechanization in problem solving: The effect of Einstellung. Psychological Monographs, 54(6).

Marsh, J. K., & Ahn, W.-K. (2006). Order effects in contingency learning: The role of task complexity. Memory & Cognition, 34(3), 568–576.

Ross, L., Lepper, M. R., & Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. Journal of Personality and Social Psychology, 32(5), 880–892.

Staw, B. M. (1976). Knee-deep in the big muddy: A study of escalating commitment to a chosen course of action. Organizational Behavior and Human Performance, 16(1), 27–44.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185(4157), 1124–1131.

Wang, Y., Cai, Y., Chen, M., Liang, Y., & Hooi, B. (2023). Primacy effect of ChatGPT. Proceedings of EMNLP 2023, 108–115.
