Why Aren’t People Calibrated on AI?

A Framework for Understanding Belief Formation About Emerging Technology

Teddy Wright & Valen Cole • University of Utah • AI + Ethics Workshop, April 2026

Companion Reference Document — Full Paper

Abstract

This paper develops a framework for understanding why people form miscalibrated views about artificial intelligence. We argue that miscalibration is not primarily an information deficit problem but an optimization target problem: individuals whose goal is identity preservation rather than accuracy deploy their cognitive capacity selectively, producing motivated reasoning that resists correction. We formalize this as a multiplicative model — View = Optimization Target × (Information Quality × Processing Quality) — and identify failures along each component, distinguishing between external factors that degrade information and processing independent of the individual and target-driven distortions that systematically bias both from within. We describe a critical property of identity-protective cognition, the self-concealing property, by which the bias is phenomenologically indistinguishable from the epistemic rigor it displaces, making standard interventions of "more awareness" insufficient. We further argue that miscalibration is compounded by misspecification of the object of the view itself — misattributing outcomes to AI rather than the human system, and treating "AI" as a single entity rather than a bundle of capabilities entering distinct domains. We then outline interventions organized around what organizations can change about the environment and what individuals can do given the limits of self-correction.

---

1. Why Calibration Matters

AI is restructuring labor markets, education, creative industries, information systems, and governance whether or not people engage with it accurately. Miscalibration has real costs in both directions. Those who overestimate AI capabilities make bad decisions: over-reliance, hype-driven policy, poor investments. Those who underestimate or dismiss AI miss genuine opportunities and fail to prepare for genuine disruptions. The fundamental problem is not ignorance about AI per se, but a mismatch between one's mental model and the actual system being navigated — a mismatch with downstream effects on personal, institutional, and policy decisions. Research on technology adoption consistently shows that inaccurate mental models of a technology predict misuse, underuse, and poor policy responses independent of the technology's actual capability (Gentner & Stevens, 1983; Norman, 1983).

Part of what makes calibration so difficult here is the stakes. AI is not just another technology entering existing structures and leaving them mostly intact. It is pressing on the foundations —

what makes humans feel uniquely important and meaningful: what makes human labor valuable, what education is actually developing, how creative work derives meaning, how expertise is recognized, and more. People are picking up on this, even when they can't articulate it precisely. The resistance and anxiety actually has quite a bit of rational grounds — they are a signal that something about the ground beneath familiar and seemingly untouchable structures is shifting. That is genuinely unsettling, and taking it seriously is not the same as being miscalibrated. The miscalibration enters when that legitimate unease gets channeled into motivated dismissal or uncritical hype rather than honest engagement with what is actually changing and what is not.

A terminological note: "AI" as a term covers a wide range of systems, from rule-based classifiers to robotics to generative models. Throughout this paper, we are primarily referring to large language models (LLMs) and the generative AI ecosystem built on top of them — the systems that have driven the current wave of public attention, disruption, and debate. The broader term "AI" appears where the point generalizes, but the specific phenomenon we are modeling is the public response to LLM-era capabilities.

---

2. A Model for View Formation

We propose the following model as the core mechanism generating miscalibrated AI views.

2a. The Causal Chain

The causal chain (Figure 1) is defined as:

» AI disrupts people's predictive models of how the world works

» Updating those models is costly (identity, livelihood, wellbeing)

» Because updating is costly, people optimize for stability over accuracy

» This optimization distorts both the information they seek and how they process what gets through

» The result is systematically miscalibrated views

Figure 1. The causal chain generating miscalibrated AI views. AI challenges predictive models entangled with identity, triggering an optimization for stability that systematically distorts both information seeking and evaluation.

The centrality of "predictive models" here is deliberate. People navigate the world using internal models that predict how value is created, how learning works, how expertise is recognized, and how information flows. AI challenges these models in most major life domains simultaneously. When a model is challenged, the rational response is to update it—but updating has real costs. Research on belief updating shows that people integrate desirable information (information that confirms existing beliefs) into their beliefs at higher rates than undesirable information (Sharot & Garrett, 2016; Garrett et al., 2014). When the challenged belief is entangled with professional identity, social standing, or sense of competence, the resistance intensifies further (Kunda, 1990; Pronin, Lin, & Ross, 2002).

The result is a cognitive system working to preserve the predictive model rather than update it. The causal chain above describes why this happens. The next step is formalizing what is happening — which components of belief formation are interacting.

2b. View Formation Equation

We formalize the above as: ***View** = **Optimization Target** × (**Information Quality** × **Processing Quality**)*

The equation is multiplicative, not additive. Bad information processed well still yields bad views. Good information processed badly also yields bad views. And critically, if the optimization target is not accuracy—if someone is optimizing for identity stability, social belonging, or cognitive ease—both inputs get systematically distorted regardless of their independent quality. The optimization target acts as a gating function on the entire system. This aligns with dual-process accounts of cognition: System 1 processing favors fast, low-effort conclusions that preserve existing mental models, while System 2 capacity is often recruited after the fact to rationalize those conclusions rather than to correct them (Kahneman, 2011).

The sections that follow unpack this equation. Section 3 takes the right side — information quality and processing quality — and examines what specifically degrades each, including a meta-epistemic problem that makes the degradation uniquely hard to fix. Section 4 takes the left side — the view — and examines what people are actually forming views *about*, and how

misspecifying the object of the view guarantees miscalibration before the processing even begins.

---

3. How Miscalibration Happens

The optimization target sits upstream of everything. But even downstream of a compromised target, the equation has two components that interact. Information quality is the input side and processing quality is the perception side. Both can fail independently, and the multiplicative structure means failure in either one is sufficient to compromise the view.

The natural response to this is: can't people just notice when their processing is off? This is the intuition behind calls for "more self-awareness" or "critical thinking" as interventions. And it is not wrong. The ability to audit your own belief-formation process is a real lever. But it runs into a problem that sits above both information and processing quality and represents another dimension of the calibration challenge: the very audit mechanism that can be an intervention is also subject to identity-protective cognition.

What follows unpacks each of these — the optimization target, information quality, processing quality, and the self-concealing property — and why addressing them requires structural interventions rather than individual awareness.

3a. Optimization Target

The optimization target sits upstream of the equation. It determines how someone will engage with information and process it. It can cause a person with excellent epistemic capacity to deploy it selectively if their goal is based on something other than accuracy.

Research on identity-protective cognition (IPC) provides documentation of this mechanism. Kahan (2013) demonstrated that subjects who scored highest on measures of cognitive sophistication showed the strongest evidence of identity-driven motivated reasoning—not the least. The finding is counterintuitive but robust: higher cognitive capacity, when recruited in service of identity preservation rather than accuracy, produces more sophisticated confabulations rather than better beliefs. (Kahan, Peters, Dawson, & Slovic, 2017).

Common non-accuracy optimization targets include:

- **Identity coherence**: updating would require revising who I am or what my work means.
- **Social belonging**: updating would put me at odds with my community or professional group.
- **Cognitive ease**: updating requires sustained effort and tolerance of uncertainty.

These targets are not themselves the biases and errors seen below, but they are the primary cause for their presence and persistence. Someone whose optimization target is accuracy will self-correct processing errors over time. Someone whose target is identity protection will

systematically recruit their view formation capacity in service of the wrong goal (Kunda, 1990; Mercier & Sperber, 2011).

3b. Information Quality

With the optimization target as the gating function, the next question is what happens downstream — how information quality and processing quality each break down under its influence and independent of it.

How reliable the information someone encounters about AI is depends on two categories of factors: those external to the individual and those driven by the optimization target.

External Factors

Incentive misalignment. Content creators, companies, and influencers have financial and status incentives to exaggerate AI capabilities or amplify fear. What drives engagement is not what is accurate. A tech CEO demoing a product has every reason to show the most impressive output and none to show the median one. A journalist covering AI risk gets more clicks from "AI will replace 80% of jobs" than from "AI will restructure some tasks within some occupations." The information people encounter is pre-filtered by someone else's optimization target before it ever reaches them.

Information environment degradation. Social media algorithms optimize for engagement rather than accuracy, systematically prioritizing content that is emotionally charged or controversial regardless of its truth value (Brady et al., 2023). Misinformation on platforms like X/Twitter spreads faster and reaches more users than accurate content, in part because novel and surprising claims generate more engagement (Vosoughi, Roy, & Aral, 2018). Research modeling the dynamics of popularity-based ranking algorithms on social media platforms shows a fundamental tradeoff: weighting content by social interactions like likes and shares increases engagement but also increases misinformation and polarization (Germano, Gómez, & Sobbrio, 2026).. The result is an environment where the volume of information about AI enlarges and accurate information gets diluted. Even someone with an accuracy-oriented optimization target faces a degraded signal.

Optimization Target Effects

Selective information search. When the optimization target is identity preservation, people preferentially seek out information that confirms their existing model and avoid information that challenges it. It is a filtering function on what information is allowed in (Nickerson, 1998). An educator who feels threatened by AI will gravitate toward articles about AI-generated plagiarism and students losing critical thinking skills. A tech optimist will gravitate toward capability demos and productivity gains. Both are encountering real information, it's just that the selection is driven by what confirms the model they need to protect, not by what would most accurately represent the landscape.

Asymmetric information weighting. Even when disconfirming information does get through the search filter, it is treated as less credible than confirming information independent of its actual quality (Nickerson, 1998). A professor reading a study showing that AI-assisted students performed better on a reasoning task might scrutinize the methodology, question the sample size, and wonder about confounds. That same professor reading a study showing AI-assisted students performed worse would accept the result with far less skepticism. The

information is weighted differently not because of the difference in information quality, but because one threatens the predictive model and the other supports it.

3c. Processing Quality (The Perception Side)

Even when good information makes it through, processing it accurately is a separate challenge — and it fails for reasons both external to and driven by the optimization target.

External Factors

Cognitive cost of tracking. Accurately tracking AI (or any complex topic) is genuinely hard. The technology changes fast, the capability frontier shifts in ways that break existing heuristics, and keeping up demands sustained attention and effort. A small business owner trying to evaluate whether AI could help their workflow has to sort through contradictory claims, rapidly shifting capabilities, and product marketing dressed up as technical assessment, all while running a business. Even with the best intentions, the cognitive cost of staying calibrated is high enough that many people rationally default to whatever simplified model they already hold.

Reasoning skill gap. Beyond the cost of tracking, processing AI information well requires a kind of reasoning skill that is not evenly distributed and not commonly trained. Evaluating probability claims, distinguishing correlation from causation in AI impact studies, tracking the difference between benchmark performance and real-world deployment, weighing competing expert predictions and more are all skills that require practice. Most people have not been trained in probabilistic reasoning or technology assessment. A person can be intelligent, educated, and well-intentioned and still lack the specific inferential tools needed to evaluate whether a given AI claim is credible.

Epistemic capacity gap. Beyond general reasoning skill, AI breaks the heuristics people use to evaluate technology because it disrupts the link between process and output that most frameworks depend on. When a person sees AI-generated text that reads like a competent professional wrote it, their existing model says "this required years of training and domain expertise." When they learn it took thirty seconds and a prompt, nothing in their framework explains how that is possible. The functional ability is real in that AI can produce outputs that are useful, sound, and appropriate, but what that *means* is hard to grasp because it sits in an uncomfortable space. It is clearly doing something that resembles what humans do, but it is doing it through information processing on patterns rather than through lived experience and accumulated understanding. That makes attribution genuinely strange: if the output is good, what produced the quality? If effort is no longer a reliable proxy for value, what is? If writing quality no longer signals the thinking behind it, how do you evaluate the person? These are questions the existing epistemic frameworks were never built to answer. A hiring manager who has spent twenty years evaluating candidates based on writing samples has no model for what it means when writing quality stops being a reliable signal of the person behind it. An educator grading essays has no framework for distinguishing a student who understood the material and used AI to articulate it from a student who understood nothing and used AI to bypass the learning entirely.

Optimization Target Effects

Prior distortion. In Bayesian terms, belief formation starts with a prior — how likely someone already considers a given hypothesis before encountering new evidence. A compromised optimization target anchors that prior not to evidence but to whatever position protects the existing predictive model. A tenured professor whose career is built on a skill set AI can approximate starts with a prior that AI is overhyped because the alternative is threatening to their professional identity. A venture capitalist whose portfolio depends on an AI boom starts with a prior that AI is transformative. Both priors are doing identity work, not epistemic work.

Likelihood distortion. The second element in Bayesian belief formation is the likelihood — how much weight new evidence gets given the hypothesis. A compromised optimization target evaluates this asymmetrically. Evidence consistent with the existing model is treated as highly probable and reliable; evidence that contradicts it is treated as less likely, methodologically suspect, or anecdotal (Lord, Ross, & Lepper, 1979; Kunda, 1990). A graphic designer who sees AI struggle with hands in an image generation tool treats that as strong evidence that AI art is fundamentally limited. That same designer seeing AI produce a stunning, coherent piece treats it as an anomaly or a cherry-picked demo. The evidence is evaluated differently not because it differs in quality but because the likelihood function is running in service of a conclusion the designer needs to reach. This is the mechanism documented in Kahan et al. (2017): higher gave people better tools for constructing arguments in defense of whatever conclusion their identity demanded.

3d. The Self-Concealing Property

The external factors described above — incentive misalignment, information environment degradation, cognitive cost, reasoning skill gaps, epistemic capacity gaps — are problems that could in principle be addressed by giving people better information and better tools for processing it. The target-driven distortions are different. Selective search, asymmetric weighting, prior distortion, likelihood distortion are problems of misdirected capacity. The natural next thought is: if people could just become more aware of when their optimization target is compromising their information intake and processing, they could catch it and correct. This intuition calls for critical thinking and epistemic self-awareness: develop the meta-skill of auditing your own belief-formation process, and you can self-correct across domains.

The intuition is not wrong in principle. The ability to notice which optimization function you are running, whether your information inputs are adequate, and whether your processing is calibrated for the question at hand is a real capacity. When it works, it is high-leverage — it enables correction across domains. But it runs into an issue because of the same optimization target that drives the other distortions: the very cognitive process one would use to detect identity-protective reasoning is itself compromised by the same identity-protective optimization target. We call this the self-concealing property (Figure 2).

What happens is when a person's optimization target shifts from accuracy to identity preservation, the shift does not produce a felt experience of "I am now being defensive." It produces a felt experience of "I am being appropriately skeptical." Identity-protective cognition

does not announce itself as bias. These two states — critical evaluation and identity-protective dismissal — feel identical from the inside. A professor who dismisses emerging AI capability data is not thinking "I am protecting my professional identity." They are thinking "I have seen hype cycles before, and I am applying well-earned skepticism." Introspection cannot distinguish these, because introspection is downstream of the optimization target, not upstream of it.

Figure 2. The self-concealing property of identity-protective cognition. The cycle is invisible to the agent running it: introspection operates downstream of the compromised optimization target, making protective processing phenomenologically indistinguishable from genuine rigor.

The bias blind spot literature provides empirical grounding for this: individuals consistently rate themselves as less subject to cognitive biases than their peers because the introspective access that would reveal the bias is precisely what the bias inhibits. (Pronin, Lin, & Ross, 2002).

3e. Consequences for Intervention

The framework so far identifies four dimensions along which views about AI break down: the optimization target that gates the entire system, information quality, processing quality, and the self-concealing property.

The external factors are addressable through external means: better information sources, better tools, better training. These are real problems and the interventions are real, but they are also the easier part of the challenge. A person whose optimization target is accuracy will benefit from better information and better processing tools. They will seek them out.

The harder problem is that the target-driven distortions — the ones that systematically bias which information gets in and how it gets processed — cannot be reliably detected or corrected by the person running them. This is what the self-concealing property establishes. And it has a counterintuitive implication: intelligent, epistemically sophisticated people are not less

vulnerable to these distortions. They may be more vulnerable, because higher cognitive capacity, when recruited in service of identity preservation, produces more sophisticated rationalizations rather than better beliefs (Kahan et al., 2017). The professor has better arguments for dismissal than the layperson does. The tech executive has a more articulate case for optimism. The sophistication of the defense scales with the sophistication of the defender.

This means the intervention framework cannot rely primarily on individual awareness, reflection, or self-correction. The external problems need external solutions, but the target-driven problems also need external solutions — mechanisms that operate on the belief-formation process from outside, because asking the compromised system to audit itself is precisely what the self-concealing property predicts will fail. Section 5 outlines what those structural interventions look like.

---

4. What Views Are Actually About

Section 3 described how views break down — the mechanisms by which optimization targets, information quality, processing quality, and the self-concealing property distort belief formation. But there is a separate problem upstream of all of that: people are often miscalibrated about what they are forming views about in the first place. Two forms of this stand out — misattributing outcomes to the wrong causal agent, and treating "AI" as a single object when it is not.

4a. The Misattribution Problem

AI is fundamentally input-output: it processes what it receives and produces outputs accordingly. The outputs are largely a function of the user's input including both context and intentions. While AI has its own effect due to its training and architecture, much of the panic and much of the hype around AI gets attributed to the technology itself rather than to the human system it enters.

A student who uses AI to skip learning was probably already disengaged. A person harmed by an AI companion was probably already vulnerable. An artist threatened by AI generation was already in a field where the bottleneck was shifting away from technical execution. AI does not have zero independent effect and it often requires more skill or awareness to use well, but treating it as the primary causal agent rather than as an amplifier leads to responses that address the wrong variable.

This happens in both directions. People misattribute bad outcomes to AI when the real source is the user, the context, or the system the AI entered. And people misattribute good outcomes to AI when the real driver was the quality of the person's input, judgment, or problem selection. Both forms produce miscalibrated views.

Where AI should be the object of evaluation is narrower than most discourse treats it. It makes sense to evaluate what the technology enables as functional capabilities — what it can and cannot do, where its outputs are reliable and where they are biased, where it performs well and where it breaks down. Critiques of AI bias, hallucination rates, and failure modes in specific deployment contexts are about the technology. But the question of whether AI is "good" or "bad" for education, creativity, or any domain is almost always a question about the human system.

4b. The Bundling Problem

"AI" is not one thing. It is a bundle of capabilities entering distinct domains, and each combination produces different effects. Treating AI as a single entity to hold one opinion about is itself a form of miscalibration. A person who says "AI is overhyped" may be right about one capability-domain pair and completely wrong about another.

These are the key domains where AI is pressing on predictive models:

Work and Labor. AI can perform tasks that previously required years of specialized training. It unbundles jobs into component tasks, some of which it can do and some of which it cannot.

Education and Learning. AI can produce assignment-quality work across most academic domains. It can also tutor, explain, and adapt to individual learners in real time.

Creative Production. AI can generate images, music, text, and video at near-zero marginal cost. It can iterate on creative work faster than humans can and in response to natural language direction.

Relationships and Social Connection. AI can hold context across long conversations, remember what someone has shared, and respond with attentiveness and availability that do not depend on another person's capacity or schedule.

Information and Epistemics. AI can synthesize large volumes of information, surface patterns across sources, and generate summaries and analyses at a speed and scale humans cannot match. It can also generate convincing but false or biased content at the same speed and scale.

But even this decomposition is insufficient. Each domain contains sub-questions that require further breakdown. AI in education means something different for formative assessment than for professional credentialing. AI in creative production means something different for graphic design than for literary fiction. AI in labor means something different for entry-level task

execution than for senior strategic judgment. Some of what AI enables in each domain is genuinely new capability. Some is a shift that is different but neutral by changing how something is done without making it better or worse. Some is different and genuinely worse for the people affected. And much of what people fear is possible but not inevitable, and is contingent on choices about deployment, policy, and institutional design that have not yet been made. Forming a calibrated view requires engaging at this level of specificity, which is precisely what a bundled framing prevents.

The interventions that follow in Section 5 are designed with this full picture in mind — addressing not just the mechanism failures from Section 3, but also the definition-level confusions that make accurate views difficult even when the mechanisms are working.

---

5. Implications for Intervention

In this section, we divide the intervention space along a line: what organizations and institutions can do, and what individuals can do.

5a. What Organizations Can Do

Improve the information landscape. The external information environment is degraded by incentive misalignment and algorithmic amplification of engagement over accuracy. Organizations producing or curating AI information: media outlets, research institutions, professional associations, and platforms, can raise the floor by separating capability reporting from marketing, requiring confidence levels on claims, and prioritizing accuracy of signal over other things. This is not a novel intervention and is the same problem science journalism has faced for decades, applied to a domain moving fast.

Standardize how AI is reported and evaluated. Any institutional communication about "AI" that does not specify which capability, in which domain, for which users, in which context, is contributing to miscalibration rather than reducing it. Organizations can push toward reporting standards that require this specificity. A headline that says "AI threatens jobs" is not well packaged information. A report that says "LLM-based code generation tools reduced time-to-completion for junior developers on routine tasks by 40% but showed no improvement for senior developers on architectural decisions" is. Professional associations, policy bodies, and media organizations can enforce frameworks that require this level of decomposition, making it harder for clumped claims to circulate unchecked.

Create structured spaces for disagreement. If the self-concealing property means individuals cannot reliably detect their own identity-protective reasoning, someone with a different identity investment may be able to. Cross-disciplinary evaluation pairs, adversarial collaboration protocols, and structured disagreement exercises create conditions where the

blind spots of one agent are visible to another. This must be built into institutional process, not left as optional. Telling people to "seek out disagreement" fails because the self-concealing property makes their current level of agreement feel earned rather than protective (Mercier & Sperber, 2011).

Invest in education. This is the broadest lever and it operates on three fronts.

First, reasoning skills. The reasoning skill gap is real and produces miscalibration independent of the optimization target. Training in probabilistic thinking, distinguishing correlation from causation, evaluating competing predictions under uncertainty, and calibrating confidence to evidence are all general epistemic tools that improve processing quality across domains. A person who can think in base rates and likelihoods evaluates AI impact claims differently than a person relying on anecdote and gut feel (Kahneman, 2011). This is not AI-specific education. It is education in how to think about uncertain, complex, rapidly changing systems — a skill set that has always been undersupplied relative to its importance.

Second, what AI actually is and how to view it. The epistemic capacity gap described in Section 3c exists because people lack mental models for what these systems do. This means engaging honestly with the hard questions AI raises rather than flattening them into simple narratives. It also means accurate AI literacy. This also includes teaching people to distinguish what should be attributed to AI from what should be attributed to the user and context. As discussed in Section 4a, much of what gets credited or blamed on AI is actually a function of who is using it and how. Training people to ask "what is the technology actually doing here, and what is the human system around it doing" is a basic literacy move that reduces misattribution in both directions. And this especially includes reality based education about what AI is. These systems are performing billions of computations to find statistical structure in massive amounts of human text, refined through extensive human feedback, producing something that captures real patterns in the world of language and knowledge even if the mechanism is unlike how humans think. Understanding the intricacies and implications here gives people better heuristics for evaluation. So does understanding that what comes out is heavily shaped by what goes in, and that capability is not uniform across domains and tasks. AI is a technology that works in a fundamentally different way than human cognition, and evaluating it requires frameworks built for that rather than borrowed from how we evaluate people.

Third, teaching the mechanisms described in this paper. Making identity-protective cognition visible as a concept — that it exists, that it scales with cognitive sophistication, that it feels like rigor rather than bias — does not eliminate the problem. The self-concealing property means awareness alone is insufficient. But it lowers the threshold for recognizing the pattern after the fact, creates shared language for naming it in groups, and provides a framework for understanding why smart people disagree so sharply about AI (Kahan et al., 2017; Pronin et al., 2002). Teaching people that their resistance to updating may be driven by what the update

would cost them — not by the quality of the evidence — is not a cure, but it is a starting condition for the structural interventions above to work.

5b. What Individuals Can Do

Individual interventions are inherently limited by the self-concealing property — you cannot fully audit a system from inside it. But partial self-correction is possible and worth pursuing.

Evaluate against outside criteria. When assessing AI outputs or AI-related claims, measure against defined external standards rather than gut reaction. The question is not "does this feel right" but "does this meet the criteria." If your evaluation of a piece of work changes when you learn AI was involved, the evaluation is measuring your comfort, not the work's quality.

Seek blind comparison. Where possible, evaluate AI-produced and human-produced work without knowing which is which. This strips the identity trigger from the assessment and lets processing normalize. This is the logic of blind peer review applied to a new domain.

Use other people's eyes. Actively solicit evaluation from people with different identity investments than your own. A colleague in a different field, a person with a different relationship to AI, someone whose professional identity is not threatened or validated by the same claims yours is. Their blind spots will differ from yours in ways that make the comparison informative.

Interrogate comfort. Most epistemic hygiene focuses on questioning beliefs we disagree with. The self-concealing property suggests the opposite: question the beliefs that feel right. When you encounter an AI take that confirms your existing model — when the reaction is "I knew it" — that is precisely the moment to apply scrutiny. Not because the conclusion is necessarily wrong, but because the feeling of rightness is exactly what identity-protective processing produces.

5c. A Note on AI as a Tool for Its Own Navigation

There is an irony in the calibration problem. The technology disrupting predictive models is also the most powerful tool available for rebuilding them. AI can synthesize information about its own impact across domains, model scenarios, surface tradeoffs, and help people think through what specific capabilities mean for their specific context. The worries people have about AI are often valid — but navigating them well frequently requires getting closer to the problem rather than further from it, and AI can help with that. Resistance to using the tool to understand the tool is itself a product of the miscalibration this paper describes.

---

6. Conclusion

Miscalibration about AI is not a deficit of information. It is a consequence of how AI disrupts predictive models that are entangled with identity, livelihood, and wellbeing. When updating is costly, people optimize for stability rather than accuracy, and they do so through a mechanism that feels indistinguishable from genuine epistemic rigor.

The framework presented here — View = Optimization Target × (Information Quality × Processing Quality) — is intended as a diagnostic tool. Each component fails in distinct ways and requires distinct interventions. External factors like information environment degradation, cognitive cost, and reasoning skill gaps are addressable through better information, better tools, and better training. Target-driven distortions like selective search, asymmetric weighting, prior distortion, and likelihood distortion require structural interventions that operate from outside the individual's compromised epistemic process. The self-concealing property explains why the second category cannot be left to self-correction, and why intelligence and sophistication do not protect against it.

The problem is further compounded at the level of what people are forming views about. Misattributing outcomes to AI rather than the human system it enters, and treating AI as a monolith rather than decomposing it into specific capabilities in specific domains, produce miscalibration before any processing even begins.

The interventions we outline are not exhaustive, but they share a common structure: they do not ask compromised systems to audit themselves. Organizations can improve the information landscape, standardize how AI is reported, create structured spaces for disagreement, and invest in education that builds reasoning capacity, AI literacy, and awareness of the mechanisms described here. Individuals can evaluate against outside criteria, seek blind comparison, use other people's perspectives, and develop the habit of questioning the views that feel most comfortable. Neither set of interventions is sufficient alone.

Getting this right matters. AI is not a temporary disruption that will stabilize into familiar patterns. The questions it raises about labor, education, creativity, relationships, and information are not going away, and the quality of the answers depends on the calibration of the people and institutions producing them. The technology that is disrupting our predictive models is also the most powerful tool available for rebuilding them — but only if we can engage with it accurately enough to use it well.

References

Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2023). How social media algorithms exploit human biases to amplify partisan content. Trends in Cognitive Sciences, 27(8), 731–744.

Garrett, N., Sharot, T., Faulkner, P., Korn, C. W., Roiser, J. P., & Dolan, R. J. (2014). Losing the rose tinted glasses: Neural substrates of unbiased belief updating in depression. Frontiers in Human Neuroscience, 8, 639.

Gentner, D., & Stevens, A. L. (Eds.). (1983). Mental models. Lawrence Erlbaum Associates.

Germano, F., & Sobbrio, F. (2026). Ranking for engagement: How social media algorithms fuel misinformation and polarization. Journal of Public Economics, 241, 105267.

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. Judgment and Decision Making, 8(4), 407–424.

Kahan, D. M., Peters, E., Dawson, E., & Slovic, P. (2017). Motivated numeracy and enlightened self-government. Behavioural Public Policy, 1(1), 54–86.

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

Kunda, Z. (1990). The case for motivated reasoning. Psychological Bulletin, 108(3), 480–498.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. Journal of Personality and Social Psychology, 37(11), 2098–2109.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. Behavioral and Brain Sciences, 34(2), 57–74.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology, 2(2), 175–220.

Norman, D. A. (1983). Some observations on mental models. In D. Gentner & A. L. Stevens (Eds.), Mental models (pp. 7–14). Lawrence Erlbaum Associates.

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. Personality and Social Psychology Bulletin, 28(3), 369–381.

Sharot, T., & Garrett, N. (2016). Forming beliefs: Why valence matters. Trends in Cognitive Sciences, 20(1), 25–33.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151.