The Illusion of Sufficiency: Can Knowing the Names of Cognitive Biases Cure the Ignorance of Our Age?

Second article in the "Illusion of Sufficiency" trilogy

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, the second in the "Illusion of Sufficiency" trilogy, tests a common belief: that knowing the name of a cognitive bias is half the battle in correcting it. Drawing on Wilson and Brekke's (1994) four-condition framework and behavioural evidence from anchoring experiments, it shows that awareness of a bias, even when a person is explicitly warned about it, does not on its own correct the resulting error in judgement. This article calls that gap "meta-ignorance". It then shows that meta-ignorance does not stay at the individual level: the bias label becomes a weapon for discrediting rivals in identity conflicts, and the attention economy of the media strips it of its empirical and methodological grounding, as the fate of the Dunning-Kruger effect in public discourse shows. The article's claim, however, is corrective rather than refuting. The available meta-analytic evidence shows that training which includes practice, feedback and real application can produce a small but meaningful reduction in bias, though its transfer to real-world decisions remains uncertain, and decision-environment design offers an independent route that can complement or substitute it. The overall conclusion is that knowing a name builds awareness but does not correct behaviour; real correction comes only through practice and feedback.

Key words: Cognitive biases; bias blind spot; meta-ignorance; debiasing training; Dunning-Kruger effect; identity-based labelling; decision-environment design; critical thinking

---

Introduction and the Place of This Article

This article is the second in a trilogy under the shared title "The Illusion of Sufficiency". The first article showed that recognising "logical fallacies", however useful, is not on its own enough to counter the ignorance of our age. This article asks the same question of a different content area: "cognitive biases". Fallacies come from the tradition of logic, and biases from the tradition of cognitive psychology. Methodologically, both fall under the broader umbrella of "critical thinking". In other words, these are two content areas to which critical-thinking skills are applied, not skills in their own right.

The third article in this series will address a gap recognised within the critical-thinking literature itself, between having the skill and actually being disposed to use it.

Before continuing, however, another boundary needs clarifying: the boundary between "cognitive biases", the subject of this article, and what is sometimes grouped, in everyday language and even in some sources, under the broader label "cognitive errors".

The general label "cognitive error", in practice, often covers lists drawn from an entirely separate tradition: that of "cognitive distortions", which belongs to the field of cognitive-behavioural therapy. Beck (1976) first formulated this idea conceptually. Burns then turned it into a precisely enumerated list of ten items (Burns, 1980), later revising it in a 1999 update (Burns, 1999).

Over time, however, later popular sources and applications have often offered a broader, inconsistent list, including items such as mind reading, catastrophising, "shoulds", personalisation, emotional reasoning, and the like (Sage, Keppens & Yannakoudakis, 2025). This list resembles "cognitive bias" in name, but it is of a different kind: a clinical tool for identifying thought patterns linked to depression and anxiety in therapy clients, not a research finding about errors of judgement in ordinary people. This article deliberately does not enter that field, not because it considers it unimportant, but because mixing it with cognitive biases under one shared label repeats exactly the confusion that Section 3 of this article criticises in another case: the detachment of a scientific label from its precise context and research tradition. This article's subject is specifically cognitive biases in the precise sense, meaning the judgement-and-decision-making research tradition, the tradition we will later see shaped by Kahneman, Tversky, and those who continued their work.

Now that the scope of this work and its conceptual boundaries are clear, we can turn to the article's subject and its purpose in the present day.

Writing an article on various aspects of cognitive biases seems justified because, in recent years, knowing the names of dozens of examples of what the public and critical-thinking enthusiasts call "cognitive biases" has become one of the most in-demand genres of public content. A clear example of this demand is Buster Benson's "Cognitive Bias Cheat Sheet" (Benson, 2016), which John Manoogian III soon turned into a visual map titled the "Cognitive Bias Codex". This map organises more than 180 biases into four categories and has since been widely reshared and used in educational and business settings (Manoogian, 2016).

In fact, the popular belief that "knowing the name of an error pattern is half the way to overcoming it" is now common enough to have earned its own name. Laurie Santos and Tamar Gendler, both of Yale University, call it the "G.I. Joe Fallacy" (Santos & Gendler, 2014), referring to the tagline of the 1980s G.I. Joe cartoon, which ended each episode with "knowing is half the battle." Kristal and Santos later examined this phenomenon directly, in 2021, under the name "G.I. Joe phenomena" (Kristal & Santos, 2021). Their explanation is that metacognitive awareness of a bias, on its own, does little to reduce it in practice.

This article puts exactly this belief and definition to the test for cognitive biases. Its precise thesis is as follows:

Conceptual knowledge about a bias (its name, definition and examples) is not the same as the capacity to detect and correct it in the moment of action. The available evidence shows that the transfer from the former to the latter, without practice, feedback and sometimes a change in the decision environment, is limited and unreliable. This claim rejects neither of two things: it does not claim that such knowledge is useless, nor that no training works at all. What it rejects is the belief that merely knowing the name is, by itself, sufficient.

Rather than offering an exhaustive catalogue of cognitive biases, this article examines a small set of representative cases (anchoring, the bias blind spot, and the Dunning-Kruger effect) as case studies for testing a broader claim about the category as a whole.

At this point, however, a question needs addressing in advance: what relation does a cognitive bias, as an error of judgement, bear to ignorance in its general sense, the subject of this series?

This article's answer lies in behavioural evidence, not in people's own reports of the "bias blind spot" (the common belief that we see ourselves as less biased than others). We know that even when a person is told explicitly that they are subject to a specific bias, this awareness alone does not correct their judgement (Section 1). And when training goes beyond merely knowing the name, the extent to which it transfers to real decision-making remains uncertain and unreliable (Section 4).

On the basis of these two documented behavioural findings, discussed in detail below, this article argues that a person may be fully aware of a limitation in their own judgement and still be unable to correct it in the moment of action. This gap between conceptual awareness and behavioural correction is what this article calls "meta-ignorance": not a lack of knowledge about the external world, but a failure of knowledge to translate into action. This thesis does not rest solely on any one of the competing explanations for the mechanism behind the blind spot (Section 1).

This meta-ignorance, not knowing one's own inability to correct a bias, does not remain confined to the individual, cognitive level. The very label that was meant to be a tool for self-knowledge becomes, in the field of identity conflict, a weapon against rivals, and in the media's attention economy, it is stripped of its scientific weight and reduced to a shallow accusation. We turn to the evidence on where and to what extent real training works only after that.

To examine these issues, the article is structured as follows:

Section 1: the theoretical framework, that is, why knowing the name of a bias is not enough to correct it;

Section 2: how the bias label turns into a weapon in identity conflict;

Section 3: how the label becomes detached from its scientific meaning in the attention economy and the media;

Section 4: the evidence showing when, and to what extent, real training, unlike merely knowing the name, is effective.

---

1. Theoretical Framework: Why Knowing the Name Is Not Enough

Meta-ignorance, the gap we outlined between knowing and correcting, needs a more precise formulation. We need to state exactly which conditions must be met for a biased judgement to be corrected, and exactly which of them knowing the name of a bias fails to meet.

On this point, Wilson and Brekke (1994) identify four conditions necessary for correcting any biased judgement: awareness that a bias exists; motivation to correct it; knowledge of the direction and size of its effect; and the ability to control or apply a corrective strategy.

Of this list, knowing the name of a bias meets, at best, only the first condition. The other three, motivation, knowledge of the size and direction of the error at that particular moment, and the capacity to apply a strategy, neither follow from knowing the name nor are guaranteed by it.

For this claim to be testable, two terms need defining precisely from the outset.

First, "knowing the name", in this article, means mere exposure to the name, definition and examples of a bias. It is something gained by reading a text, seeing an infographic, or hearing a talk, without active practice, corrective feedback, or the experience of applying it in a real decision.

Second, "training" means any other kind of exposure that includes at least one of three elements: practice, feedback, or application in a simulated or real situation.

This distinction matters because, if a study were to show that mere exposure to the name and definition of a bias, without any of those three elements, reduces real judgement error, this article's thesis would be refuted.

Evidence supporting this point comes from Wilson and colleagues' 1996 study (Wilson, Houston, Etling & Brekke, 1996). In this study, participants were given an anchor number, drawn from their own identification number, and asked to compare it with the number of physicians listed in their city's telephone directory, then give their own estimate. Before this comparison, one group of participants was told explicitly that a number already in their mind could affect their answer to the following question. This fully and directly met the first condition of Wilson and Brekke's model: awareness that a bias exists. The estimates of all anchored groups, warned and unwarned alike, still differed significantly ($p < .05$) from those of the unanchored control group, and no significant difference emerged between the warned and unwarned anchored groups themselves. In other words, the warning alone neither eliminated the anchoring effect nor even weakened it slightly.

From a cognitive-psychology perspective, this gap can also be seen more precisely within Kahneman's two-system framework (Kahneman, 2011). "Knowing the name of a bias" is a product of System 2: slow, deliberate, and effortful. The occurrence of the bias itself, however, takes place in System 1: fast and automatic, before System 2 gets the chance to intervene. From this perspective, even when System 2 knows the name of the bias, its inherent laziness means that, in the moment of action, it usually yields the floor to System 1.

Rozenblit and Keil showed a parallel example of this gap under the name "Illusion of Explanatory Depth" (Rozenblit & Keil, 2002). In their main experiment, people believed they understood the mechanisms of simple everyday devices, from toilet siphons to cylinder locks, until asked to explain them step by step. At that point, the depth of their understanding turned out to be far shallower than they had assumed.

In another example from the same line of research, this time concerning complex skills, Lawson (2006) tested this for how bicycles work, and Fernbach and colleagues (Fernbach, Rogers, Fox & Sloman, 2013) tested it for policy positions. Both reached similar results. By analogy, one might infer that knowing the name of a bias creates an illusion of the same kind: a person believes that, because they know its name, they have mastered it, without ever having tested that mastery in a real situation.

Here, it should be made clear that this step is an analogy, not a conclusion drawn directly from the studies above, whose subject was other skills, not cognitive biases themselves. The shared mechanism in both cases, however, is the same: an illusion of mastery, arising from mere exposure, without practical testing.

Even that first, simple condition, awareness that a bias exists, is not always easily met, let alone its effect on correcting our judgements. Pronin, Lin and Ross showed this, for example, when they asked participants to rate their own degree of bias and that of an average American across several well-known biases (Pronin, Lin & Ross, 2002). People consistently rated themselves as less biased

than others. The authors named this pattern the "bias blind spot". Their original explanation, in that paper, was that a person's own perception of reality feels unmediated and objective, so anyone who disagrees with that perception must be either mistaken or biased.

A few years later, Pronin and colleagues formulated a different explanation, the "introspection illusion" (Pronin, Gilovich & Ross, 2004; Pronin & Kugler, 2007). On this account, introspection, reviewing one's own thoughts and intentions, is a reliable test for assessing one's own bias, whereas for others we have access only to their outward behaviour. Since bias usually operates unconsciously, introspection does not reveal it in ourselves, even though behaviour reveals it in others.

Neither of these two explanations, however, can be treated as final today, since further work has been done in this area. Mandel, Collins, Walker, Fugelsang and Risko, for instance, tested two further explanations directly, alongside the introspection-illusion account (Mandel, Collins, Walker, Fugelsang & Risko, 2022).

One is the "cognitive sophistication" explanation. This holds that the blind spot is not really an error at all, but a reflection of reality: people with better cognitive ability and thinking dispositions really are less biased than others, so their lower self-ratings are accurate, not mistaken.

The other is the "conversational processes" explanation, which treats the blind spot as a product of how the question is framed, not a genuine error in self-knowledge. On this view, when a participant is told that a given bias is common among people in general and is then asked to rate the average person, they conclude, following ordinary conversational norms, that this generalisation is about others, not necessarily about themselves.

Mandel and colleagues found, in this study, the following. The conversational-processes explanation received the most support, because most participants rated others above the midpoint of the scale, while their self-ratings fluctuated more and rarely reached that same point. The cognitive-sophistication explanation was not supported, because, on objective bias tasks, people's actual level of bias was unrelated to their level of cognitive sophistication. The introspection-illusion explanation was likewise not supported, because the pattern it predicted, relying more on introspection when judging oneself and more on behavioural tendencies when judging others, did not appear in the data. In addition, the introspection-illusion score correlated only weakly with the overall size of the blind spot, and, after controlling for cognitive ability and the disposition toward actively open-minded thinking (AOT), it was no longer a significant predictor.

There is a subtle point here that needs clarifying.

If the self-other asymmetry is, to some extent, a product of how the question is framed, as the conversational-processes explanation suggests, this might seem to call the whole phenomenon of the blind spot into question, and with it this article's reliance on it.

This is not the case. The dispute over the blind spot's mechanism, why this asymmetry occurs, is independent of the behavioural evidence this article's thesis relies on. West, Meserve and Stanovich showed, in their own research, that higher cognitive ability does not reduce the blind spot, and in some analyses even slightly increases it (West, Meserve & Stanovich, 2012). Scopelliti and colleagues likewise showed that people with a larger blind spot accept others' advice less and perform worse on objective decision-making tasks (Scopelliti et al., 2015).

These two findings were obtained independently of any of the three competing explanations. They show that a person, regardless of which mechanism underlies their blind spot, still cannot correct

it simply by knowing its name. In other words, the dispute over mechanism gives us a more precise picture of the blind spot, without changing the behavioural conclusion this article's thesis rests on.

Apart from the dispute over mechanism, something should also be said about the phenomenon of the bias blind spot itself.

This same self-other asymmetry, of which the bias blind spot is one specific case, has in fact been reproduced fairly consistently in later research. One preregistered replication in Brazil, for example, confirmed the effect in a population outside the usual North American student samples (Seda et al., 2024). This same replication, however, also revealed a limitation in the original theory: contrary to Pronin and colleagues' prediction, the self-other asymmetry was not confined to bias; it also appeared for more general personal shortcomings.

Another replication, with samples from Hong Kong and the United States, confirmed the same asymmetry in a different design (Chandrashekar et al., 2021), this time also linking it to people's beliefs about free will. A cautious conclusion one can draw from this research is that the phenomenon itself is on firmer ground than its theoretical explanations, and today's ongoing dispute concerns why it occurs, not whether it exists.

An important point to bear in mind here is that this description of the blind spot does not mean permanent, incurable blindness. In Section 4 we will see that, under specific training conditions, that is, as defined above, training that includes practice and feedback, not merely knowing the name, the blind spot itself can be reduced. What this section shows is limited to the claim that merely knowing the name of a bias, without such training, does not reduce the blind spot.

There is, admittedly, also a limitation within this framework itself that should be noted: Gigerenzer's (2018) warning about what he calls "bias bias". This is the tendency to see every deviation from an ideal standard as a cognitive error, meaning that even detecting bias can itself be biased, when many of these patterns are not flaws at all but heuristics suited to, and effective within, their real environment. Gigerenzer calls this concept "ecological rationality".

This criticism does not invalidate this article's thesis. Even if some of these patterns are adaptive heuristics rather than errors, turning their names into commodities, without understanding the conditions of their application, remains misleading. The criticism is, however, a reminder that the category of cognitive bias itself is not unified or free of dispute.

The picture so far is still incomplete. What we have seen describes the bias blind spot mainly as a fairly neutral cognitive limitation, an error in self-assessment that every person, regardless of their social or political position, is caught in to roughly the same degree. But the next question is this: is this asymmetry itself distributed equally among people and groups in practice? Or does its pattern of occurrence and use also follow the same social and political divides it was supposed merely to make us aware of?

We turn to this question in the next section.

---

2. The Bias Label as an Identity Weapon

If the blind spot were simply a neutral cognitive limitation, its pattern of occurrence should also be distributed equally among everyone. There is evidence, however, showing that this asymmetry is itself distributed in a biased way.

Walker, Collins, Walker, Fugelsang and Mandel, in a paper in press and published online in May 2026, showed across three experiments that the more positively a person feels toward someone,

the less bias they attribute to them. It is this affective evaluation that determines the amount of bias attributed. In partisan experiments, for instance, Democrats and Republicans attributed far more bias to their partisan opponents than to members of their own party. This finding is a more recent version of what Ehrlinger, Gilovich and Ross had shown earlier, in 2005: outgroup attribution of bias (Ehrlinger, Gilovich & Ross, 2005).

Kennedy and Pronin's study closes this research loop. They showed that seeing the other side as biased itself intensifies conflict (Kennedy & Pronin, 2008). This means the bias label is not merely a consequence of partisan hostility but also reinforces it. In such conditions, attributing bias to others can turn from a documented critique of a reasoning process into person-centred discrediting, much as the social psychologist Lee Jussim, in an unreviewed piece, describes the same pattern for confirmation bias (Jussim, 2019). In his view, without specific evidence of selective use of data or disregard for contrary evidence, the accusation of confirmation bias itself is nothing more than a disguised personal attack.

This identity function, however, depends on one precondition: that the name of the bias can be detached from its precise empirical grounding and turned into a simple, widely understood tool, ready for use in argument. The next section shows a close example of exactly this process of detachment.

---

3. The Attention Economy: When a Label Detaches from Its Empirical Grounding

As we have seen, the bias label acquires, in identity conflict, a function detached from its empirical content. Once the name of a bias is detached from its empirical context and methodological conditions, it becomes a commodity that spreads more easily than the skill of detection itself. A clear example of this process is the fate of the Dunning-Kruger effect.

Technically, this effect is, first and foremost, a metacognitive phenomenon, a person's perception of their own ability, not a perceptual or memory bias of the anchoring or confirmation type. This difference in kind is often ignored in the term's public and media use.

Krueger and Mueller first raised systematically, in 2002, the idea that the Dunning-Kruger pattern might be partly a product of regression to the mean and the better-than-average effect in self-assessment (Krueger & Mueller, 2002). Gignac and Zajenkowski then added a narrower, more methodological critique in 2020 (Gignac & Zajenkowski, 2020). In their view, proving a real, non-linear effect requires a significant quadratic term in polynomial regression; otherwise, the pattern could simply reflect a linear relationship.

This is a methodological critique, not a wholesale denial of the phenomenon, just as Jansen, Rafferty and Griffiths have argued, from a different angle, that part of the real effect remains (Jansen, Rafferty & Griffiths, 2021).

What matters for this article, beyond the details of this dispute, is that the popularised label, "less able people are more confident", has drifted far beyond the precise scope and the point of contention in the original study. It is exactly this simplified label that is also used in political discourse: "the other side is ignorant and therefore confident", a pattern that aligns with the identity function of the confirmation-bias label discussed in the previous section.

This pattern, a label that extends beyond its precise scope and becomes a tool for argument, might wrongly suggest that training people about biases is essentially useless. This is not the case. The next section shows that properly designed training can genuinely have an effect, though with its own limits.

---

4. The Counter-Evidence: When Training Works, and Its Limits

This article's claim is a limited one: knowing the name alone is not enough, not that no training has any effect at all. Morewedge, Yoon, Scopelliti, Symborski, Korris and Kassam showed, in 2015, that a single interactive training session measurably reduced bias up to eight weeks later, in Study 1, and twelve weeks later, in Study 2 (Morewedge et al., 2015). Non-interactive video training, which explained the definition, example and strategy rather than merely the name, also had a significant and fairly lasting effect, around a 19 to 26 per cent reduction, though smaller than that of the interactive game. In Study 1, the bias blind spot itself was one of the three biases trained, and it showed a reduction too.

It should be noted, for the sake of transparency, that two of this paper's authors, Symborski and Korris, were affiliated with Creative Technologies Inc., the company that made the "Missing" training game. The research was designed under the IARPA programme named SIRIUS and funded under a contract from the US Air Force Research Laboratory. The paper nonetheless carries an official statement of no conflict of interest. This point is raised here purely for transparency, not to cast doubt on the results.

The next field study of this kind, conducted in 2019, reproduced this same training effect among 290 students across three professional programmes at HEC Paris, using a case based on the Challenger shuttle-launch decision (Sellier, Scopelliti & Morewedge, 2019). Trained individuals chose the weak, confirmation-seeking solution 19 per cent less often than untrained individuals.

To generalise these results to a larger scale, we turn here to two complementary reviews.

The first is the work of Swaryandini and colleagues, with Noetel as senior author (Swaryandini et al., 2025). In 2025, they published a meta-analysis of 41 trials, drawing on 160 effect sizes out of 383 identified across 54 trials, with 10,941 participants in total, on the effectiveness of bias-reduction training among school and university students. According to their results, the overall effect, based on the effect-size measure Hedges' g, a statistic devised by the statistician Larry Hedges that, like Cohen's d, measures the size of an effect in a meta-analysis, was small but significant: 0.26, with a confidence interval of 0.14 to 0.39.

All these studies carried an unclear or high methodological risk, signs of publication bias were also found, and the authors themselves wrote explicitly that it remains unclear how far these improvements transfer to real decision-making outside the classroom.

The second study is by Fasolo, Heard and Scopelliti, who, in the same year, reviewed one hundred studies of organisational decision-making from a different angle (Fasolo, Heard & Scopelliti, 2025). They examined 32 studies of individual-level training interventions, 62 studies of decision-environment design, and 6 studies combining both. Their conclusion was that no study has directly compared the two approaches, and which is superior remains unknown.

Even so, in their view, the choice between the two approaches depends on context. Individual correction, that is, training, appears more suitable for complex, poorly structured decisions, where enough time and attention are available. Decision-environment design is the better option for simpler, more stable decisions, where people's goals are aligned or staff turnover is high.

Together, these two reviews give a more precise picture, made up of three independent points. Bias training can genuinely have an effect, but its effect size is usually small. On methodological quality, the literature in this field is not uniform, and its transfer to real decision-making remains

an open question. On decision-environment design, as a third, independent route alongside individual correction, no research has directly compared the two approaches, but each appears more effective in its own domain: training is more accessible and more scalable for complex, poorly structured decisions, while decision-environment design, though it requires a choice architect and separate adjustment for each context, works more effectively for simpler, more stable decisions.

This section's conclusion is corrective to the article's thesis, not a refutation of it. What works is not the name alone, but training that includes definition, example, practice and feedback, and even then its effect size is small and its transfer to the real world remains uncertain. Decision-environment design, as a third, independent route, can serve as a complementary or alternative option for some biases.

In this way, the very meta-ignorance with which this piece began, not knowing that merely knowing the name of a bias is not enough to correct it, is, at least for the reader of these lines, no longer meta-ignorance.

In the end, from all of this, perhaps one final conclusion can be drawn. A name builds awareness, but merely knowing the name of a bias is, for correcting it, about as useful as knowing the name of a disease is for curing it: almost nothing. A bias cannot be defeated by naming it; it can only be resisted through practice and persistence.

Conclusion

This article, the second link in the "Illusion of Sufficiency" trilogy, pursued a limited but well-documented claim: knowing the name of a cognitive bias is a necessary condition for correcting it, not a sufficient one.

Section 1 showed that this gap is not merely a theoretical guess. It can be formulated precisely within the Wilson and Brekke framework, and it has been repeatedly documented behaviourally, in anchoring experiments and in the bias-blind-spot literature. Even where the exact mechanism behind the blind spot remains disputed, the phenomenon itself is not.

Sections 2 and 3 then showed that this meta-ignorance does not stay neutral in the real world. The very label meant to serve as a tool for self-knowledge becomes, in identity conflicts, a weapon against rivals, and in the media's attention economy, it is detached from its precise methodological grounding and reduced to a simple slogan, a fate of which the Dunning-Kruger effect is a clear example.

Section 4, however, showed that this picture is not one-sided. When training goes beyond merely knowing the name and includes practice, feedback and real application, it produces a real, measurable effect. That effect, according to the available meta-analytic evidence, is small, the methodological quality of the literature is not uniform, and its transfer to real decision-making remains uncertain. Alongside training, decision-environment design is also put forward as an independent, sometimes more effective route, though no research has yet directly compared the two approaches.

Taken together, these four sections neither deny the value of knowing the names of biases nor prescribe training unconditionally. Rather, they draw a precise line between what knowing can achieve and what it cannot. This line is, at the same time, the point where this article connects to the rest of the trilogy. If the first article showed that knowing fallacies alone is not enough, and this article has shown that knowing biases alone is not enough either, the third article addresses a

different gap of the same kind: between having the skill of critical thinking and actually being disposed to use it. Meta-ignorance, then, is not a feature of one particular field of knowledge, but a recurring pattern in how human beings relate to knowledge about their own limitations.

References

Beck, A. T. (1976). Cognitive therapy and the emotional disorders. International Universities Press.

Benson, B. (2016). Cognitive bias cheat sheet. Better Humans (Medium).

Burns, D. D. (1980). Feeling good: The new mood therapy. William Morrow.

Burns, D. D. (1999). Feeling good: The new mood therapy (Rev. ed.). Avon Books.

Chandrashekar, S. P., Yeung, S. K., Yau, K. C., Cheung, C. Y., Agarwal, T. K., Wong, C. Y. J., Pillai, T., Thirlwell, T. N., Leung, W. N., Tse, C., Li, Y. T., Cheng, B. L., Chan, H. Y. C., & Feldman, G. (2021). Agency and self-other asymmetries in perceived bias and shortcomings: Replications of the bias blind spot and extensions linking to free will beliefs. Judgment and Decision Making, 16(6), 1392–1412.

Ehrlinger, J., Gilovich, T., & Ross, L. (2005). Peering into the bias blind spot. Personality and Social Psychology Bulletin, 31(5), 680–692.

Fasolo, B., Heard, C., & Scopelliti, I. (2025). Mitigating cognitive bias to improve organizational decisions. Journal of Management, 51(6), 2182–2211.

Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. Psychological Science, 24(6), 939–946.

Gigerenzer, G. (2018). The bias bias in behavioral economics. Review of Behavioral Economics, 5(3–4), 303–336.

Gignac, G. E., & Zajenkowski, M. (2020). The Dunning-Kruger effect is (mostly) a statistical artefact. Intelligence, 80.

Jansen, R. A., Rafferty, A. N., & Griffiths, T. L. (2021). A rational model of the Dunning-Kruger effect supports insensitivity to evidence in low performers. Nature Human Behaviour, 5, 756–763.

Jussim, L. (2019, August 7). Confirmation bias: Real bias or delegitimization rhetoric? Psychology Today ("Rabble Rouser" blog). https://www.psychologytoday.com/us/blog/rabble-rouser/201908/confirmation-bias-real-bias-or-delegitimization-rhetoric

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

Kennedy, K. A., & Pronin, E. (2008). When disagreement gets ugly: Perceptions of bias and the escalation of conflict. Personality and Social Psychology Bulletin, 34(6), 833–848.

Kristal, A. S., & Santos, L. R. (2021). G.I. Joe phenomena: Understanding the limits of metacognitive awareness on debiasing (Working Paper No. 21-084). Harvard Business School.

Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. Journal of Personality and Social Psychology, 82(2), 180–188.

Lawson, R. (2006). The science of cycology: Failures to understand how everyday objects work. Memory & Cognition, 34(8), 1667–1675.

Mandel, D. R., Collins, R. N., Walker, A. C., Fugelsang, J. A., & Risko, E. F. (2022). Hypothesized drivers of the bias blind spot: Cognitive sophistication, introspection bias, and conversational processes. Judgment and Decision Making, 17(6), 1392–1421.

Manoogian, J., III. (2016). The cognitive bias codex [Infographic].

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing decisions: Improved decision making with a single training intervention. Policy Insights from the Behavioral and Brain Sciences, 2(1), 129–140.

Pronin, E., Gilovich, T., & Ross, L. (2004). Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others. Psychological Review, 111(3), 781–799.

Pronin, E., & Kugler, M. B. (2007). Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. Journal of Experimental Social Psychology, 43(4), 565–578.

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.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. Cognitive Science, 26(5), 521–562.

Sage, A., Keppens, J., & Yannakoudakis, H. (2025). A survey of cognitive distortion detection and classification in NLP (arXiv:2508.09878). arXiv.

Santos, L. R., & Gendler, T. S. (2014). [Essay on the "G.I. Joe fallacy"]. In J. Brockman (Ed.), This idea must die: Scientific theories that are blocking progress. Harper Perennial. (As cited in Kristal & Santos, 2021.)

Scopelliti, I., Morewedge, C. K., McCormick, E., Min, H. L., Lebrecht, S., & Kassam, K. S. (2015). Bias blind spot: Structure, measurement, and consequences. Management Science, 61(10), 2468–2486.

Seda, L., Martins, I. T. C., Lisboa, T. L. R. C., Batistuzzo, M. C., & Fatori, D. (2024). Theoretical maturation of the "bias blind spot": A preregistered replication study of Pronin, Lin, and Ross (2002) in a Brazilian sample. Collabra: Psychology, 10(1), Article 122158. https://doi.org/10.1525/collabra.122158

Sellier, A.-L., Scopelliti, I., & Morewedge, C. K. (2019). Debiasing training improves decision making in the field. Psychological Science, 30(9), 1371–1379.

Swaryandini, G., Graham, J., Griffith, S., Grilo, V., Ruzzante, F., Zhang, X., Yeung, S. K., Mangiarulo, M., Basarkod, G., Ng, C., Parker, P., Tangen, J., Saeri, A., Grundy, E., Slattery, P., & Noetel, M. (2025). Systematic review and meta-analysis of educational approaches to reduce cognitive biases among students. Nature Human Behaviour, 9(12), 2510–2538.

Walker, A. C., Collins, R. N., Walker, H. E. K., Fugelsang, J. A., & Mandel, D. R. (2026). Everyone I don't like is biased: Affective evaluations and the bias blind spot. Personality and Social Psychology Bulletin. Advance online publication.

West, R. F., Meserve, R. J., & Stanovich, K. E. (2012). Cognitive sophistication does not attenuate the bias blind spot. Journal of Personality and Social Psychology, 103(3), 506–519.

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgment and evaluation. Psychological Bulletin, 116(1), 117–142.

Wilson, T. D., Houston, C. E., Etling, K. M., & Brekke, N. (1996). A new look at anchoring effects: Basic anchoring and its antecedents. Journal of Experimental Psychology: General, 125(4), 387–402.
