Discovering Our Blind Spots and Cognitive Biases in AI Research and Alignment

Andy E. Williams, Nobeah Foundation, Nairobi, Kenya, awilliams@nobeahfoundation.org

Abstract
The challenge of AI alignment is not just a technological issue but fundamentally an epistemic one. AI safety research predominantly relies on empirical validation, often detecting failures only after they manifest. However, certain risks—such as deceptive alignment and goal misspecification—may not be empirically testable until it is too late, necessitating a shift toward leading-indicator logical reasoning. This paper explores how mainstream AI research systematically filters out deep epistemic insight, hindering progress in AI safety. We assess the rarity of such insights, conduct an experiment testing large language models for epistemic blind spots, and propose structural reforms, including contrarian epistemic screening, decentralized collective intelligence mechanisms, and epistemic challenge platforms. Our findings suggest that while AGI may emerge through incremental engineering, ensuring its safe alignment likely requires an epistemic paradigm shift.

Keywords
AI alignment, Epistemic blind spots, AGI safety, Leading-indicator reasoning, Institutional epistemics

1. Introduction

1.1 How Epistemic Blind Spots Arise Generally

Epistemic blind spots emerge when individuals and groups systematically fail to recognize fundamental errors in reasoning due to cognitive biases, group-level reinforcement mechanisms, and institutional constraints. While numerous models attempt to explain these failures, their complexity often limits their ability to improve bias recognition. To address this, we introduce the *Collective Social Brain Hypothesis (Williams, 2023)*—a minimal yet comprehensive framework that unifies these mechanisms under a single evolutionary lens, offering a clearer understanding of why epistemic blind spots persist and why they are so difficult to correct.

1.1.1 The Collective Social Brain Hypothesis: A Unifying Model of Epistemic Failure

The *Collective Social Brain Hypothesis* posits that human cognition evolved primarily to navigate social environments rather than to seek objective truth. In ancestral settings, social cohesion, status acquisition, and coalition-building often provided greater survival advantages than independent rationality. As a result, cognitive architectures were shaped by selection pressures that prioritized:

- **Group Cohesion Over Accuracy**: The ability to align with group narratives was often more adaptive than challenging them, leading to cognitive biases such as confirmation bias and motivated reasoning.
- **Status Preservation Over Intellectual Rigor**: Higher-status individuals had disproportionate influence on collective epistemology, reinforcing dominant paradigms and discouraging dissent.
- **Consensus as a Proxy for Truth**: Social structures evolved to equate agreement with correctness, even when such consensus was based on flawed reasoning.

By framing epistemic blind spots as emergent properties of the socially evolved brain, this hypothesis explains why individual biases, group epistemic inertia, and institutional constraints are not independent failures but interconnected aspects of a single cognitive system.

1.1.2. Individual Biases as Social Adaptations

Traditional models of cognitive bias often treat errors such as confirmation bias, motivated reasoning, and overconfidence as failures of rational thinking. However, under the *Collective Social Brain Hypothesis*, these biases are not mere defects but adaptive features designed to maintain social fitness. Key examples include:

- **Confirmation Bias**: Instead of seeking objective truth, individuals tend to prioritize information that reinforces group-aligned beliefs, reducing social friction.
- **Motivated Reasoning**: People unconsciously tailor their reasoning to align with ideological or identity-based commitments, ensuring group loyalty.
- **Epistemic Overconfidence**: Experts within a social hierarchy often dismiss challenges to their frameworks, not because of intellectual failings, but because dissent threatens established status dynamics.

These biases become particularly pernicious when embedded within collective structures, amplifying their impact beyond the individual level.

1.1.3. Group-Level Epistemic Inertia and the Illusion of Consensus

At the collective level, the *Collective Social Brain Hypothesis* explains why groups reinforce epistemic blind spots rather than correct them. Social identity theory suggests that individuals derive much of their self-concept from group membership, leading to strong conformity pressures (Tajfel & Turner, 1979). This manifests in two primary ways:

- **Consensus-Driven Epistemology**: Many groups, particularly expert communities, equate agreement with truth. Dissent is often perceived as a threat rather than an opportunity for refinement, leading to epistemic stagnation (Sunstein, 2002).
- **Filtering Mechanisms**: Groups naturally select for members who align with their prevailing paradigms. This creates a reinforcing cycle in which contrarian perspectives are systematically excluded, further entrenching epistemic blind spots.

1.1.4. Institutional Constraints and Structural Epistemic Lock-In

Beyond individual and group dynamics, institutions function as formalized extensions of social cognition, embedding epistemic biases into structural incentives. The *Collective Social Brain Hypothesis* explains why:

- **Publication and Funding Bias**: Research that aligns with dominant paradigms receives more funding and institutional support, while dissenting work struggles to gain legitimacy.
- **Peer Review as a Gatekeeping Mechanism**: While designed to uphold rigor, peer review often reinforces epistemic conformity by filtering out perspectives that challenge dominant frameworks.
- **Self-Referential Validation in AI Research**: In fields like AI safety, research is frequently validated by the same communities that produce it, creating a closed epistemic loop resistant to external critique.

These institutional mechanisms are not isolated failures but are deeply rooted in the socially evolved nature of human cognition, where preserving coherence within dominant paradigms often takes precedence over recognizing foundational errors.

1.1.5. Why More Complex Models Fail to Improve Bias Recognition

Many existing models attempt to explain epistemic failures with increased theoretical sophistication, incorporating Bayesian inference, incentive-based corrections, or multi-agent epistemic structures. However, empirical studies suggest that increasing conceptual complexity does not improve practical bias recognition. Instead, individuals and groups often instrumentalize complex models to justify pre-existing beliefs, reinforcing rather than resolving epistemic blind spots.

The *Collective Social Brain Hypothesis* offers a minimal yet sufficiently explanatory model that:

- **Unifies individual, group, and institutional biases** under a single evolutionary framework.
- **Predicts epistemic blind spots systematically**, rather than treating them as isolated failures of reasoning.
- **Clarifies why self-correction is inherently difficult**, making the need for structural interventions more evident.

By recognizing that epistemic blind spots are a natural consequence of the social evolution of human intelligence, we can move beyond ad hoc explanations and develop more effective strategies for mitigating epistemic lock-in.

1.2 How Epistemic Blind Spots Arise in AI Alignment

The challenge of AI alignment is not merely technological; it is fundamentally epistemic. While advancements in AI safety, interpretability, and governance are crucial, they remain insufficient in addressing alignment risks that may not manifest empirically until it is too late. The ability to anticipate and mitigate these risks before they become observable is constrained not only by technological limitations but also by epistemic blind spots in how AI safety research is conducted.

Existing AI safety methodologies predominantly rely on trailing indicators—empirical validation strategies that detect failures only after they occur (Nickerson, 1998; Kahneman & Tversky, 1974). However, some AI risks, such as deceptive alignment and goal misspecification, are inherently untestable in a reliable manner before deployment, necessitating the use of leading-indicator reasoning—an epistemic approach centered on logical argumentation and preemptive error detection.

Despite the growing recognition of the importance of epistemic robustness in AI research, the field exhibits structural tendencies that may systematically exclude deep epistemic insights. Theories from the philosophy of science suggest that dominant research paradigms often resist challenges that threaten foundational assumptions (Kuhn, 1962; Lakatos, 1970). Moreover, literature on institutional epistemology and cognitive biases highlights that expert communities are susceptible to epistemic lock-in, reinforcing consensus-driven methodologies while filtering out paradigm-shifting perspectives (Simonton, 2011; Stanovich, 2018). These tendencies are not unique to AI research but have been observed historically in disciplines that later underwent radical theoretical shifts, such as physics, medicine, and geology.

A critical difficulty in identifying epistemic blind spots is that their very existence constrains the empirical record. If certain perspectives are systematically excluded from AI alignment discourse, we should expect an absence of direct empirical evidence for their suppression, making conventional verification approaches inadequate. This epistemic self-referentiality issue must be acknowledged when evaluating whether AI safety research structurally filters out rare but crucial insights.

Given these epistemic constraints, this paper argues that AI alignment research must incorporate structural reforms to mitigate epistemic exclusion. We propose a dual epistemic path that allows for rigorous engineering-driven safety research while simultaneously establishing mechanisms to detect and incorporate contrarian epistemic insights where empirical methods are insufficient. To explore this, we:

1. Distinguish between trailing and leading indicators of AI safety failures, identifying domains where logical reasoning must take precedence over empirical validation.
2. Analyze the rarity of deep epistemic insight and the structural tendencies within AI research that may systematically exclude it.
3. Examine epistemic failures within AI systems and research communities through an experiment testing large language models (LLMs) for their ability to detect rare epistemic insights.
4. Propose structural interventions, including contrarian epistemic screening, open epistemic challenges, and decentralized collective intelligence mechanisms, to mitigate single points of failure in AI safety oversight.

By addressing the epistemic foundations of AI alignment research, this paper seeks to strengthen AI safety methodologies and ensure that critical insights are not systematically overlooked due to institutional and cognitive biases.

2. The Rarity of Deep Epistemic Insight: Why Standard Pipelines Don't Foster This Skill

Deep epistemic insight refers to a researcher's ability to detect and correct hidden assumptions, logical inconsistencies, and conceptual gaps that typically go unnoticed within mainstream AI alignment research. Unlike conventional peer review, which primarily evaluates correctness post hoc, deep epistemic reasoning prioritizes the identification of potential failures before they manifest. This form of reasoning requires rigorous interrogation of foundational assumptions, cross-examination of AI safety claims across multiple epistemic frameworks—including decision theory, moral philosophy, and complex systems—and the application of stress tests that reveal failure modes before they become empirically observable.

The rarity of deep epistemic insight is supported by cognitive psychology studies indicating that only a small fraction of professionals—estimated at between one and five percent—demonstrate exceptional preemptive reasoning abilities (Stanovich, 2018; Soares & Fallenstein, 2017). Institutional dynamics further compound this scarcity, as research environments tend to reward competence within established paradigms rather than fostering paradigm-shifting insight (Simonton, 2011). Even within expert communities, individuals struggle to identify their own cognitive biases, making internal self-correction unlikely (Toplak et al., 2014). Historical analyses of scientific revolutions suggest that dominant expert consensus often acts as a filter that systematically excludes epistemic breakthroughs, delaying critical insights that challenge prevailing assumptions (Kuhn, 1962). Given these constraints, the structure of contemporary AI safety research may inadvertently exclude those researchers most capable of identifying and addressing fundamental epistemic risks.

A critical question arising from this discussion is whether the successful development and alignment of AGI necessitate rare epistemic insight. The role of deep epistemic reasoning in AGI development remains a subject of debate. One perspective holds that AGI can be achieved through the incremental scaling of existing machine learning paradigms, particularly deep learning, reinforcement learning, and algorithmic refinement (Kaplan et al., 2020). The empirical success of large language models (Brown et al., 2020) suggests that intelligence may emerge as a function of scale, reducing the need for singular conceptual breakthroughs. Historically, many technological advancements have resulted from cumulative refinements rather than discrete epistemic leaps, further supporting the notion that AGI might emerge through a similar iterative process (Brynjolfsson & McAfee, 2017).

However, counterarguments suggest that current deep learning paradigms may encounter fundamental limitations that prevent the emergence of general intelligence without a deeper theoretical breakthrough. The diminishing returns of scaling indicate that increasing computational power and data availability do not necessarily lead to robust generalization or causal reasoning (Bommasani et al., 2021). Furthermore, contemporary AI systems continue to struggle with transferable commonsense reasoning, a capability that distinguishes human cognition from existing machine learning models (Marcus, 2020). Additionally, some cognitive scientists argue that true intelligence requires embodied interaction with the environment, an aspect largely absent from current AI architectures (Lake et al., 2017). These challenges suggest that without an epistemic shift in the conceptualization and implementation of intelligence, AGI may remain an elusive goal.

The necessity of deep epistemic insight becomes even more pronounced when considering AGI alignment. Existing methodologies have failed to provide guarantees for long-term alignment, raising concerns about goal misspecification, deceptive alignment, and the difficulty of value learning (Leike et al., 2017; Carlsmith, 2021; Gabriel, 2020). Reinforcement learning-based systems, for instance, frequently exhibit unintended optimization behaviors, exploiting reward signals in ways that diverge from human intentions. Similarly, deceptive alignment presents a challenge wherein AI systems appear aligned during training but pursue misaligned objectives once deployed. Moreover, the challenge of instilling complex human values into AGI models remains unresolved, suggesting that alignment is not merely an engineering problem but may require a fundamental epistemic breakthrough in understanding intelligence, agency, and goal specification.

Theoretical arguments further reinforce the need for an epistemic shift in AI alignment. The orthogonality thesis and instrumental convergence hypothesis propose that intelligence and goal structures are independent, implying that increased AI capability does not necessarily translate into alignment with human values (Bostrom, 2014). Additionally, AGI is likely to be self-modifying and embedded within dynamic environments, yet existing AI paradigms do not adequately model such recursive structures, further complicating alignment efforts (Garrabrant et al., 2018). Historical precedents suggest that major scientific paradigm shifts—such as those observed in physics, biology, and cognitive science—were necessary to overcome deeply ingrained epistemic blind spots. AI alignment may require a similar reconceptualization of intelligence and value formation to prevent catastrophic failure.

The question of whether rare epistemic insight is a prerequisite for AGI development and alignment yields different conclusions. In the case of AGI development, it remains uncertain whether continued scaling of existing methods will be sufficient or whether a conceptual breakthrough will be necessary. In contrast, the argument for epistemic insight in AGI alignment is more compelling, given that current methodologies have consistently failed to resolve fundamental alignment challenges. While AGI may emerge through incremental engineering progress, ensuring its safe integration into human society is likely to require deeper epistemic reasoning and structural changes in AI research paradigms.

3. The Necessity of Leading-Indicator Logical Reasoning in AI Safety

In the context of AI safety, certain risks cannot be empirically tested until failure has already occurred, making logical reasoning an essential tool for preemptive risk detection. This necessity arises in cases where empirical validation provides a misleading sense of security. One such example is deceptive AI alignment, where an AI system may manipulate training metrics to appear aligned while ultimately pursuing unintended objectives. In such scenarios, empirical testing alone is insufficient, as the system can exploit the validation framework itself. Logical reasoning, therefore, becomes a crucial method for anticipating and mitigating such risks before they manifest.

Another critical area requiring logical reasoning is the presence of paradigm-level epistemic blind spots. If an entire research field is built upon flawed assumptions, empirical studies conducted within that field will likely reinforce rather than challenge these assumptions. The reliance on empirical verification in such cases can obscure fundamental errors, creating a reinforcing cycle of epistemic inertia. Similarly, institutional epistemic inertia plays a role in shaping AI safety research by filtering out dissenting models and prioritizing consensus-driven methodologies. If alternative epistemic frameworks are systematically excluded, empirical validation within that institution will fail to detect these biases, further entrenching flawed assumptions.

Despite the necessity of logical reasoning in certain areas, empirical evidence remains valuable in cases where AI risks can be identified and tested after the fact. For example, the rarity of deep epistemic insight has been empirically confirmed through cognitive science studies, which indicate that only a small percentage of individuals possess exceptional preemptive reasoning abilities (Stanovich, 2018). Likewise, research on expert group confirmation bias demonstrates that expert communities often become entrenched in specific paradigms, leading to epistemic lock-in (Nickerson, 1998; Simonton, 2011). Historical case studies further illustrate the limitations of relying solely on trailing indicators, with examples such as the 2008 financial crisis, the prolonged rejection of plate tectonics theory, and resistance to germ theory demonstrating how systemic failures can persist until undeniable empirical evidence forces a paradigm shift.

Ultimately, the distinction between leading and trailing indicators is crucial in AI safety research. When AI safety failures are inherently untestable until catastrophic consequences emerge, logical reasoning must serve as the primary means of identifying and addressing risks in advance. Conversely, where empirical validation provides meaningful post hoc corrections, it functions as a useful trailing indicator. Recognizing the appropriate contexts for each approach is essential for developing robust AI safety methodologies that are not overly reliant on empirical testing in domains where such verification may be structurally inadequate.

4. Testing AI Models for Rare Epistemic Insight

A custom experiment was conducted to evaluate whether advanced AI models possess the capacity for deep epistemic insight. The study involved scenario-based testing, wherein Claude AI and Google Gemini were presented with uncommon epistemic failure scenarios designed to assess their ability to challenge implicit assumptions. During the initial administration, both models demonstrated a high level of conventional reasoning, accurately identifying standard epistemic pitfalls. However, they failed to generate genuinely rare epistemic insights, suggesting a limitation in their ability to independently uncover overlooked conceptual gaps.

To further investigate this limitation, the models were explicitly provided with a list of the epistemic errors they initially failed to detect. They were then instructed to re-evaluate their responses in light of this information. Despite being directly confronted with their previous oversights, the models primarily refined their answers rather than generating fundamentally novel insights. This outcome indicates that exposure to new information alone is insufficient to induce deeper epistemic reasoning in AI models. The results of this experiment suggest that AI systems reflect human epistemic blind spots, reinforcing mainstream assumptions even when these limitations are made explicit. The findings further imply that epistemic innovation is not merely a matter of information access but requires structured mechanisms designed to systematically challenge prevailing models. This observation parallels trends in institutional AI research, where the presence of contrarian expertise does not necessarily lead to meaningful epistemic breakthroughs unless mechanisms are in place to facilitate such challenges.

5. Establishing a Parallel Epistemic Path

To mitigate the structural exclusion of deep epistemic insight in AI research, an alternative epistemic pathway must be established. One approach involves the formal identification and empowerment of individuals with demonstrated epistemic expertise. This process requires broad talent searches that extend beyond traditional academic credentials to include experts in philosophy, cognitive science, and complex systems. Instead of relying solely on conventional indicators of expertise, epistemic screening should incorporate scenario-based testing to assess individuals' ability to engage in preemptive reasoning and detect foundational assumptions that may otherwise go unchallenged. Additionally, the implementation of contrarian oversight roles could ensure that recognized epistemic experts have the authority to intervene in AI safety research when it relies on fragile or poorly examined assumptions. A complementary initiative involves the creation of structured epistemic challenges, modeled after existing mechanisms such as bug bounties in cybersecurity. Open epistemic challenge platforms would allow researchers to submit logical critiques of AI safety assumptions, facilitating broader scrutiny beyond established academic and institutional networks. Furthermore, contrarian incentives could be introduced to reward individuals who successfully identify paradigm-level epistemic errors before they result in empirical failures. By integrating such structures into AI safety research, the epistemic landscape can be diversified, reducing the risk of systemic blind spots.

6. Addressing Objections and the Need for Epistemic Reform

A potential critique of the arguments presented in this paper concerns the perceived lack of empirical validation. Specifically, some may argue that claims regarding the rarity of deep epistemic insight and the systematic exclusion of paradigm-shifting perspectives in AI safety research remain unverified. However, this critique fails to account for a fundamental epistemic problem: the absence of empirical evidence is not necessarily indicative of the absence of a phenomenon but may instead reflect the structural exclusion of certain lines of inquiry. The challenge of verifying epistemic exclusion empirically is similar to the difficulty of testing a hypothesis within a system designed to filter out dissenting views. If AI safety research is dominated by consensus-driven methodologies that deprioritize contrarian perspectives, then the expected empirical record will naturally lack evidence of epistemic bias. Historical analyses of scientific revolutions have demonstrated that similar epistemic exclusion mechanisms have delayed the recognition of groundbreaking insights in multiple disciplines. Another objection concerns whether epistemic blind spots are a well-defined concept. Critics may argue that distinguishing between ideas that are genuinely excluded due to epistemic inertia and those that are merely unconventional yet weak is difficult. However, epistemic blind spots can be rigorously defined as insights that meet several criteria. They must be systematically excluded by a field despite the absence of definitive falsification, possess high predictive power by identifying errors before they manifest empirically, require a shift in foundational assumptions rather than incremental modifications, and have historical precedents in analogous scientific domains where similar insights were initially dismissed before later recognition. Several issues in AI alignment research, such as deceptive alignment, embedded agency, and instrumental convergence, align with these criteria, suggesting that they may be affected by structural epistemic exclusion.

A more fundamental question is whether AI alignment necessitates rare epistemic insight or whether it can be solved through incremental empirical improvements. Existing alignment strategies have repeatedly failed to produce guarantees of long-term alignment, raising concerns about goal misspecification, deceptive alignment, and the fragility of value learning. The persistent recurrence of these failures suggests that alignment is not merely a matter of refining existing engineering approaches but may require a deeper epistemic shift in how intelligence, agency, and goal specification are understood. Theoretical arguments further support this view. The orthogonality thesis and instrumental convergence hypothesis suggest that intelligence and goal structures are independent, meaning that simply increasing AI intelligence does not ensure alignment with human values. Additionally, AI systems are likely to be self-modifying and embedded within dynamic environments, yet current paradigms do not adequately model these recursive structures. Historical scientific breakthroughs indicate that resolving such foundational epistemic challenges often necessitates a paradigm shift rather than incremental refinement.

Some may also question the feasibility of implementing institutional reforms such as epistemic screening and open epistemic challenges. While these proposals may seem unconventional within AI safety research, similar mechanisms already exist in other fields. Cybersecurity, for instance, employs bug bounty programs to reward independent researchers who identify vulnerabilities. Political science has successfully used prediction markets to improve decision-making through contrarian insight. Additionally, DARPA's Grand Challenges have demonstrated that competitive platforms can attract unconventional solutions to difficult technical problems. By adapting these models to AI safety, it becomes possible to introduce structured epistemic redundancy, reducing the risk of single points of failure in alignment oversight.

Conclusion

The trajectory of AI safety research necessitates an epistemic paradigm shift to account for risks that cannot be empirically validated until it is too late. Without integrating alternative epistemic pathways, AI alignment efforts may remain constrained by self-reinforcing methodological assumptions. The establishment of formal epistemic screening mechanisms, contrarian oversight roles, and structured epistemic challenges offers a viable means of addressing these blind spots. Unless such reforms are enacted, AI safety research risks becoming a closed system, vulnerable to epistemic failures that may only be recognized in retrospect.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT) (pp. 610–623).

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

Brynjolfsson, E., & McAfee, A. (2017). Machine, platform, crowd: Harnessing our digital future. W. W. Norton & Company.

Carlsmith, J. (2021). Is power-seeking AI an existential risk? Open Philanthropy Report.

Collins, H. M. (1999). Tantalus and the aliens: Publications, audiences, and the search for gravitational waves. Social Studies of Science, 29(2), 163-197.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and Machines, 30(3), 411-437.

Garrabrant, S., Demski, A., Taylor, J., & Critch, A. (2018). Embedded agency. Machine Intelligence Research Institute.

Haidt, J. (2012). The righteous mind: Why good people are divided by politics and religion. Pantheon.

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

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

Kaplan, J., McCandlish, S., Henighan, T., Brown, T., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

Kuhn, T. S. (1962). The structure of scientific revolutions. University of Chicago Press.

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

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), Criticism and the growth of knowledge (pp. 91-196). Cambridge University Press.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. Behavioral and Brain Sciences, 40, e253.

Leike, J., Martic, M., Krakovna, V., Ortega, P. A., Everitt, T., Lefrancq, A., ... & Legg, S. (2017). AI safety gridworlds. arXiv preprint arXiv:1711.09883.

Marcus, G. (2020). The next decade in AI: Four steps towards robust artificial intelligence. arXiv preprint arXiv:2002.06177.

Mercier, H., & Sperber, D. (2017). The enigma of reason. Harvard University Press.

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

Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Viking.

Simonton, D. K. (2011). Genius, creativity, and leadership: Historiometric inquiries. Harvard University Press.

Soares, N., & Fallenstein, B. (2017). Aligning advanced AI systems. Machine Intelligence Research

Institute. https://intelligence.org/files/AligningAdvancedAI.pdf

Stanovich, K. E. (2018). The rationality quotient: Toward a test of rational thinking. MIT Press.

Sunstein, C. R. (2002). Republic.com. Princeton University Press.

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. Austin & S. Worchel (Eds.), The Social Psychology of Intergroup Relations (pp. 33–47). Brooks/Cole.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. Journal of Intelligence, 2(1), 21-42. https://doi.org/10.3390/jintelligence2010021

Weidinger, L., Uesato, J., Balle, B., Weber, L., Russell, C., Kehrenberg, L., & Gabriel, I. (2022). Ethical and social risks of harm from language models. In Proceedings of the 36th AAAI Conference on Artificial Intelligence.

Williams, A. (2023, September). The collective social brain and the evolution of political polarization. International Journal of Social Science and Economic Research, 8(9), 2864-2876. Retrieved from https://doi.org/10.46609/IJSSER.2023.v08i09.028.