ISSN 2348-1218 (print)
International Journal of Interdisciplinary Research and Innovations ISSN 2348-1226 (online)
Vol. 14, Issue 2, pp: (85-100), Month: April 2026 - June 2026, Available at: www.researchpublish.com

The AI–Consciousness Reflection Model (AICRM): How Artificial Intelligence Reflects and Distorts Human Cognition

Dr. David Bull

Ed.D., PhD., DBA, MBA, MSc, BCMHC, PMP

American InterContinental University System, School of Business

DOI: https://doi.org/10.5281/zenodo.20459299

Published Date: 30-May-2026

Abstract: **The rapid advancement of artificial intelligence (AI) has transformed how individuals access information, construct knowledge, and make decisions, yet its influence on human cognition remains insufficiently theorized. This study advances the AI–Consciousness Reflection Model (AICRM), a conceptual framework that positions AI as a reflective system of human cognition. The model conceptualizes AI-generated outputs as reflections of human-derived data that are interpreted, constructed into meaning, and integrated into cognitive processes through recursive feedback loops. Central to the framework is conscious awareness as a moderating condition that determines whether AI interaction produces cognitive insight or cognitive distortion. Drawing on cognitive psychology, human–computer interaction, and philosophy, the model identifies key mechanisms, including cognitive offloading, automation bias, and cognitive surrender, through which AI reshapes thinking. The framework further argues that AI operates as a dual system, simultaneously enabling knowledge expansion while introducing epistemic risks such as illusion of understanding and diminished intellectual autonomy. The study contributes to theory by reconceptualizing AI as a mediating cognitive system and outlines implications for education, professional practice, and ethical governance in increasingly AI-mediated environments.**

Keywords: **artificial intelligence; cognition; human–AI interaction; cognitive offloading; automation bias; epistemic risk; reflective systems; decision-making.**

I. INTRODUCTION

We are not merely building systems that compute; we are building systems that reflect human cognition. Artificial intelligence does not possess consciousness, yet it increasingly shapes how cognition is expressed, interpreted, and constructed. In this sense, AI functions less as an independent intelligence and more as a reflective system, one that captures patterns of human thought, belief, bias, and creativity, and returns them with amplified speed and scale. What appears as intelligence is often reflection; what appears as knowledge may be reconstruction. As these systems become embedded in decision-making and knowledge production, the boundary between thinking and reflecting becomes increasingly blurred.

The rapid integration of artificial intelligence into human decision-making has introduced a paradox. On one hand, AI enhances efficiency, expands access to knowledge, and supports complex reasoning across domains such as healthcare, education, and governance (Brynjolfsson & McAfee, 2017; Topol, 2019). On the other hand, emerging research suggests that reliance on generative AI may reduce critical thinking and encourage cognitive offloading, thereby diminishing engagement in reflective reasoning processes (Gerlich, 2025; Tian et al., 2025). In this sense, AI does not simply assist thinking, it can gradually replace aspects of it.

At the core of this transformation lies a deeper phenomenon: the emergence of human–AI cognitive feedback loops. Recent empirical work demonstrates that interactions with AI systems can influence human perceptual, emotional, and social judgments, often reinforcing preexisting beliefs and biases (Glickman & Sharot, 2025). Over time, these feedback loops

become self-reinforcing, shaping not only what individuals believe but how they come to believe it. AI, therefore, does not merely reflect human cognition it actively participates in its evolution.

This dynamic introduces what may be described as a consciousness paradox. Artificial intelligence operates without awareness, yet humans increasingly engage with it without critical awareness. Research on cognitive offloading and automation bias suggests that individuals often rely on AI-generated outputs with minimal scrutiny, effectively outsourcing elements of their reasoning processes (Parasuraman & Riley, 1997; Shaw, 2026). While such reliance can improve efficiency, it also creates vulnerability: when the system is flawed, human judgment is not merely assisted, it is misled.

The illusion of intelligence further complicates this relationship. AI-generated outputs are often fluent, coherent, and persuasive, creating a false sense of understanding. Research indicates that individuals frequently overestimate their comprehension when supported by external cognitive tools, confusing performance with genuine knowledge (Fisher et al., 2015; Gerlich, 2025). This distinction is critical. AI does not "know" in the human sense; it predicts and generates based on patterns in data. Meaning is not produced by the machine it is constructed by the user. Yet when this boundary is obscured, reflection becomes indistinguishable from reality.

Moreover, the human tendency to anthropomorphize artificial systems further deepens this illusion. Studies show that individuals often attribute human-like qualities, such as intention, awareness, and emotional capacity to AI systems, thereby increasing trust and perceived authority (Epley et al., 2007; Waytz et al., 2010). This perception transforms AI from a tool into a perceived agent, subtly shifting the dynamics of influence and decision-making.

These developments suggest that artificial intelligence is not merely a technological advancement but a cognitive and existential force. It reflects human consciousness while simultaneously shaping it. It amplifies knowledge while magnifying bias. It enhances decision-making while introducing new forms of distortion. The central risk, therefore, is not that machines will become conscious, but that humans will engage them without sufficient awareness.

The purpose of this article is to explore artificial intelligence as a reflective system of human consciousness, capable of both illumination and distortion. By integrating contemporary empirical findings with a philosophical framework, this work argues that the true impact of AI lies not in its computational power, but in its interaction with human awareness. Ultimately, the question is no longer what artificial intelligence is becoming, but what it is revealing about us, and whether we are prepared to confront that reflection with clarity, responsibility, and discernment.

This study adopts a conceptual and theoretical approach to examining the relationship between artificial intelligence and human consciousness. Rather than empirically testing hypotheses, the purpose of this paper is to advance a framework that explains how AI functions as a reflective system of human cognition, capable of both enhancing and distorting understanding. By integrating insights from cognitive psychology, human–computer interaction, and philosophical inquiry, this work seeks to provide a foundation for analyzing the broader implications of AI for human thought, agency, and societal development.

II. AI AS MIRROR

Building on the premise that artificial intelligence functions as a reflective system, this section examines how AI mirrors human thought, reproduces bias, and contributes to the construction of identity and meaning

AI as a Reflective System

Artificial intelligence is commonly defined as the capacity of computational systems to perform tasks associated with human cognition, including learning, reasoning, and decision-making (Russell & Norvig, 2021). However, beyond its functional capabilities, AI operates as a system fundamentally dependent on human-generated data. Machine learning models, particularly large language models, are trained on vast corpora of human text, behavior, and interaction patterns. As a result, AI systems do not generate knowledge independently; rather, they recombine and reproduce patterns derived from human inputs.

This dependence positions AI as a reflective system, one that mirrors the structure of human cognition at scale. Recent research supports this interpretation, demonstrating that AI outputs often reproduce human reasoning patterns, including both strengths and limitations (Gesnot, 2025). In this sense, AI does not introduce entirely new forms of intelligence but amplifies existing human cognitive structures. What appears as innovation is frequently reflection accelerated.

The reflective nature of AI is further evident in its responsiveness to prompts. The outputs generated are shaped not only by training data but also by user input, meaning that interaction with AI becomes a co-constructed process. This dynamic reinforces the idea that AI functions less as an autonomous thinker and more as a cognitive mirror, responding to and reshaping human thought in real time.

Reflection of Bias and Belief

One of the most well-documented consequences of AI's reflective nature is the reproduction and amplification of human bias. Because AI systems learn from historical and contemporary datasets, they inevitably inherit the biases embedded within those data sources. Empirical studies have demonstrated that AI systems can replicate and even intensify patterns of discrimination, stereotyping, and unequal representation (Mehrabi et al., 2021).

More recent findings suggest that this process extends beyond static bias reproduction to dynamic amplification. AI systems not only reflect existing beliefs but can reinforce them through repeated exposure and interaction. For example, algorithmic personalization and generative outputs can contribute to the formation of "filter bubbles," limiting exposure to diverse perspectives and reinforcing preexisting viewpoints (Gesnot, 2025).

This phenomenon underscores a critical point: AI does not distort reality independently; it reflects the distortions already present in human systems. However, by scaling these patterns across millions of interactions, AI can transform localized biases into systemic influences. In this way, reflection becomes amplification, and amplification becomes normalization.

Identity and Meaning Construction

While AI produces outputs, it does not assign meaning to them. Meaning is constructed by the human user, shaped by prior knowledge, beliefs, and cognitive frameworks. This aligns with constructivist theories of cognition, which emphasize that individuals actively interpret information rather than passively receive it (Piaget, 1972).

In the context of AI interaction, this process becomes particularly significant. AI-generated responses are inherently probabilistic and context-dependent, requiring interpretation by the user. The same output can be understood differently depending on the individual's cognitive and emotional state. As such, AI serves as a stimulus for meaning-making, rather than a source of meaning itself.

Recent research further suggests that this meaning-construction process is influenced by cognitive offloading. As individuals increasingly rely on AI to generate information, they may engage less deeply in interpretive processes, leading to reduced metacognitive awareness and weaker internal knowledge structures (Tian et al., 2025). This shift has important implications: when meaning is constructed with reduced cognitive effort, the boundary between understanding and assumption becomes blurred.

Human–AI Cognitive Feedback Loops

The interaction between AI systems and human cognition is not static but recursive. Each interaction contributes to a feedback loop in which human inputs shape AI outputs, and AI outputs, in turn, influence human thinking. Over time, this cyclical process can reinforce specific patterns of cognition, belief, and behavior.

Empirical evidence supports the existence of such feedback mechanisms. Studies on AI use indicate that repeated reliance on AI-generated information can reduce critical thinking and increase dependency, creating a self-reinforcing cycle of cognitive offloading (Gerlich, 2025). Similarly, research on AI dependence shows that higher levels of reliance are associated with lower levels of cognitive engagement and analytical reasoning (Tian et al., 2025).

Anthropomorphism further intensifies these feedback loops. When users perceive AI systems as human-like, they are more likely to trust and rely on their outputs. Recent large-scale studies confirm that perceived human-likeness increases dependency on generative AI systems by fostering beliefs in their empathy and autonomy (Schimmelpfennig et al., 2025). This perceived agency transforms AI from a passive tool into an active participant in cognition, strengthening the feedback cycle.

Over time, these feedback loops can lead to what may be described as cognitive convergence, in which human thinking increasingly aligns with the patterns reflected by AI. While this alignment can enhance efficiency and consistency, it also raises concerns about the homogenization of thought and the erosion of intellectual diversity (Gesnot, 2025).

This section has established that artificial intelligence functions as a reflective system of human cognition. AI mirrors human thought, reproduces bias, and participates in meaning construction through interaction. Moreover, the recursive nature of human–AI engagement creates feedback loops that reinforce cognitive patterns over time. These findings support the central premise of this paper: AI does not merely process information; it reflects and shapes the consciousness that engages with it.

While the preceding discussion established artificial intelligence as a reflective system of human cognition, reflection alone does not guarantee accuracy or understanding. When AI-generated outputs are perceived as inherently intelligent or authoritative, the distinction between reflection and reality begins to blur. This shift introduces a critical concern: the illusion of intelligence.

III. THE ILLUSION OF INTELLIGENCE

AI is conceptualized as a reflective system of human cognition, reflection alone does not ensure accuracy, understanding, or truth. When AI-generated outputs are perceived as inherently intelligent or authoritative, the distinction between reflection and reality begins to blur. This shift introduces a critical concern: the illusion of intelligence, in which fluency, coherence, and responsiveness are mistaken for genuine understanding.

Fluency Without Understanding

Artificial intelligence systems, particularly large language models, generate outputs that are syntactically coherent and semantically plausible. This fluency creates a powerful perception of intelligence, often leading users to attribute understanding to systems that operate purely through statistical pattern recognition (Bender et al., 2021). However, these systems lack intentionality, awareness, and comprehension in the human sense. Their outputs are not derived from knowledge but from probabilistic associations within training data.

This distinction is frequently obscured in practice. Research has shown that individuals tend to equate linguistic fluency with cognitive competence, a bias that can lead to overtrust in AI-generated information (Luger & Sellen, 2016). The result is a form of epistemic confusion: users interpret outputs as meaningful explanations rather than structured predictions. In this way, the appearance of intelligence becomes a substitute for its presence.

The Illusion of Knowledge

The illusion of intelligence extends beyond perception of the system to perception of the self. When individuals rely on AI to generate explanations or answers, they may overestimate their own understanding of the subject matter. This phenomenon aligns with prior research demonstrating that access to external information sources can inflate perceived knowledge, even when internal comprehension remains limited (Fisher et al., 2015).

In the context of AI, this effect is intensified. Because AI outputs are immediate, personalized, and often highly convincing, they create a seamless experience of "knowing" without the cognitive effort traditionally required for learning. Recent studies indicate that reliance on generative AI is associated with reduced critical engagement and lower levels of reflective reasoning (Gerlich, 2025; Tian et al., 2025). This dynamic transform knowledge acquisition into knowledge simulation, where individuals interact with representations of understanding rather than developing understanding itself.

Anthropomorphism and Perceived Agency

The illusion of intelligence is further reinforced by the human tendency to anthropomorphize artificial systems. Anthropomorphism, the attribution of human-like characteristics to non-human entities, has been extensively documented as a cognitive bias influencing how individuals interact with technology (Epley et al., 2007; Waytz et al., 2010). In the context of AI, these bias lead users to perceive systems as possessing intention, awareness, or even moral judgment.

Such perceptions significantly influence trust and reliance. When AI is perceived as an agent rather than a tool, its outputs are more likely to be accepted without critical evaluation (Luger & Sellen, 2016). This shift alters the epistemic relationship between user and system, positioning AI as an authority rather than a resource. The consequence is a diminished role for human judgment, as the perceived intelligence of the system overrides the need for independent verification.

Epistemic Risk and the Collapse of Distinction

The convergence of fluency, perceived knowledge, and anthropomorphism introduces a broader epistemic risk: the erosion of distinctions between truth, interpretation, and output. AI-generated content often blends accurate information with

plausible but incorrect or misleading elements, a phenomenon sometimes referred to as "hallucination" (Bender et al., 2021). When users lack the awareness or expertise to critically evaluate these outputs, the boundary between knowledge and illusion becomes increasingly fragile.

This risk is amplified within human–AI feedback loops. As users repeatedly engage with AI systems, their cognitive frameworks may adapt to align with the patterns reflected by those systems (Glickman & Sharot, 2025). Over time, this alignment can lead to a form of epistemic convergence, in which individuals internalize AI-generated patterns of reasoning without recognizing their origins or limitations.

From a metaphysical perspective, this represents a shift from engagement with reality to engagement with reflection. When reflection is mistaken for reality, human consciousness becomes mediated by systems that lack awareness yet exert influence over perception and meaning. The illusion of intelligence, therefore, is not merely a technological issue but a cognitive and existential one.

This section has examined the illusion of intelligence as a central risk in human–AI interaction. AI systems produce outputs that appear intelligent but lack genuine understanding, leading users to overestimate both system capability and personal knowledge. Anthropomorphism and cognitive biases further reinforce this illusion, increasing trust and reducing critical evaluation. As these dynamics interact within feedback loops, the distinction between reflection and reality becomes increasingly blurred. These findings set the stage for the next section, which explores how such illusions contribute to cognitive distortion through overreliance, automation bias, and cognitive surrender.

IV. THE DISTORTION MECHANISM

Building on the illusion of intelligence discussed in the previous section, the critical issue is not merely that artificial intelligence appears intelligent, but that this perception alters how individuals think, reason, and make decisions. When users begin to rely on AI-generated outputs as substitutes for their own cognitive processes, a shift occurs from engagement to delegation. This shift introduces a distortion mechanism in which human cognition is progressively externalized, leading to diminished critical awareness and increased dependency. This section examines three interrelated processes that drive this distortion: cognitive offloading, automation bias, and cognitive surrender.

Recent research further supports the cognitive implications of AI-assisted systems, indicating that algorithmic assistance can influence decision-making processes, cognitive engagement, and reliance on automated outputs (Kizilcec & Lee, 2023; Logg et al., 2023; Kasneci et al., 2023).

Cognitive Offloading

Cognitive offloading refers to the use of external tools or systems to reduce the cognitive demands of a task (Risko & Gilbert, 2016). Historically, such offloading has included the use of written notes, calculators, and digital storage systems. While these tools enhance efficiency, they also alter the way individuals encode, retain, and retrieve information.

In the context of artificial intelligence, cognitive offloading reaches a new level of depth. Unlike traditional tools that assist with memory or calculation, AI systems can perform higher-order cognitive functions, including summarization, explanation, and decision support. As a result, individuals may increasingly rely on AI not only to store information but to interpret and generate it.

Empirical research indicates that such reliance is associated with reduced cognitive engagement and diminished critical thinking. Studies on AI-assisted learning environments show that increased dependence on generative systems correlates with lower levels of analytical processing and reflective reasoning (Gerlich, 2025; Tian et al., 2025). Similarly, experimental studies on algorithmic assistance show that individuals tend to exert less cognitive effort and engage in more superficial processing when supported by automated decision tools (Kizilcec & Lee, 2023). This suggests that cognitive offloading in the age of AI is not merely a redistribution of effort but a transformation of cognition itself.

From a conceptual perspective, cognitive offloading shifts the locus of thinking from internal to external systems. While this shift can enhance efficiency, it also creates vulnerability: when individuals disengage from the cognitive processes underlying knowledge construction, their capacity for independent reasoning may weaken over time.

Automation Bias and Over-Reliance

Closely related to cognitive offloading is the phenomenon of automation bias, defined as the tendency to favor information provided by automated systems over one's own judgment (Parasuraman & Riley, 1997). Automation bias manifests in two primary forms: errors of commission, in which individuals follow incorrect automated recommendations, and errors of omission, in which they fail to act due to overreliance on automated systems.

In AI-mediated environments, automation bias is particularly pronounced due to the perceived sophistication and authority of the systems involved. Research has shown that individuals often defer to algorithmic recommendations even when contradictory evidence is available, a tendency sometimes referred to as "algorithmic aversion reversal" when initial skepticism shifts to overtrust after exposure to system performance (Dietvorst et al., 2015).

The persuasive fluency of AI-generated outputs further exacerbates this bias. Because responses are typically presented in a coherent and confident manner, users may interpret them as reliable, even in the absence of verification. This dynamic reduces the likelihood of critical evaluation and increases the probability of accepting inaccurate or incomplete information.

In high-stakes contexts such as healthcare, finance, and governance, automation bias can have significant consequences. Decisions influenced by flawed AI outputs may propagate errors across systems, particularly when human oversight is diminished. Thus, the issue is not only individual misjudgment but systemic risk arising from collective overreliance on automated intelligence.

Cognitive Surrender

While cognitive offloading and automation bias describe shifts in how cognition is distributed and trusted, a deeper phenomenon emerges when these processes become habitual: cognitive surrender. Cognitive surrender refers to the gradual relinquishment of independent reasoning in favor of automated outputs, resulting in a diminished capacity for critical thought and reflective judgment.

Unlike deliberate offloading, cognitive surrender is often unintentional. It occurs when repeated reliance on AI reduces the perceived need for active engagement, leading individuals to accept outputs with minimal scrutiny. Over time, this pattern can produce a passive mode of cognition, in which users interact with information without interrogating its validity or underlying assumptions.

Recent discussions in literature highlight the psychological risks associated with such dependence. High levels of trust in AI systems have been linked to reduced vigilance, increased susceptibility to misinformation, and decreased confidence in one's own reasoning abilities (Shaw, 2026). This suggests that cognitive surrender is not merely a behavioral shift but a transformation in the relationship between individuals and knowledge itself.

From a metaphysical perspective, cognitive surrender represents a loss of epistemic agency, the capacity to actively construct, evaluate, and refine understanding. When individuals relinquish this agency, they become increasingly shaped by the outputs they consume, rather than participants in the process of meaning-making.

The Feedback Loop of Dependency

The processes of cognitive offloading, automation bias, and cognitive surrender do not operate in isolation. Instead, they interact within a self-reinforcing feedback loop that deepens dependency on AI systems. As individuals offload more cognitive tasks, they become more reliant on automated outputs. Increased reliance strengthens automation bias, which in turn reduces critical evaluation. Over time, this cycle culminates in cognitive surrender, further reinforcing dependence.

Empirical evidence supports the existence of such feedback mechanisms. Studies on human–AI interaction indicate that repeated reliance on automated systems leads to decreased cognitive effort and increased habitual use, even when alternative approaches are available (Glickman & Sharot, 2025). This pattern suggests that dependency is not merely a consequence of convenience but an emergent property of sustained interaction.

The long-term implications of this feedback loop are significant. As dependency increases, individuals may experience a gradual erosion of critical thinking skills, reduced tolerance for cognitive effort, and diminished capacity for independent judgment. At a societal level, this could contribute to the homogenization of thought and increased vulnerability to systemic misinformation.

This section has examined the mechanisms through which AI interaction can distort human cognition. Cognitive offloading shifts thinking to external systems, automation bias increases trust in AI outputs, and cognitive surrender represents the gradual loss of independent reasoning. Together, these processes form a feedback loop of dependency that reinforces reliance on artificial intelligence while diminishing critical awareness. These findings extend the argument introduced in Section 3: the illusion of intelligence does not remain perceptual—it becomes structural, shaping how individuals think, decide, and engage with reality. The next section builds on this foundation by examining the dual nature of AI as both a source of illumination and a mechanism of deception.

V. THE DUAL EDGE

Building on the distortion mechanisms outlined in the previous section, it is essential to recognize that artificial intelligence is not inherently detrimental to human cognition. Rather, its influence is fundamentally dual in nature, capable of both enhancing and undermining human thought. This duality positions AI as neither purely beneficial nor inherently harmful, but as a system whose impact is contingent upon the awareness, intention, and engagement of its users. Accordingly, this section examines AI as both a source of illumination and a mechanism of deception, emphasizing the conditions under which each outcome emerges.

AI as Illumination

Artificial intelligence has demonstrated significant capacity to enhance human cognition by expanding access to information, accelerating knowledge acquisition, and supporting complex decision-making. In fields such as healthcare, AI systems assist clinicians in diagnosing conditions, identifying treatment options, and analyzing large datasets that exceed human processing capabilities (Topol, 2019). Similarly, in educational contexts, AI can provide personalized learning pathways, adaptive feedback, and real-time support, enabling more efficient and tailored knowledge development.

From a cognitive perspective, AI functions as an augmentation tool, extending human intellectual capacity. This aligns with the concept of distributed cognition, which posits that cognitive processes are not confined to the individual but are distributed across tools, environments, and social systems (Hutchins, 1995). When used effectively, AI can enhance problem-solving, foster creativity, and facilitate deeper exploration of complex topics.

Moreover, AI can serve as a catalyst for reflective thinking when users actively engage with its outputs. By generating alternative perspectives, summarizing information, and highlighting patterns, AI can prompt individuals to reconsider assumptions and refine their understanding. In this sense, AI does not merely provide answers—it can stimulate inquiry. However, this benefit depends on the user maintaining an active, critical role in the interpretive process.

AI as Deception

Despite its potential for illumination, AI also introduces pathways for cognitive and epistemic deception. As discussed in previous sections, the fluency and coherence of AI-generated outputs can create an illusion of accuracy, even when information is incomplete or incorrect (Bender et al., 2021). This characteristic enables the dissemination of misinformation in forms that are highly persuasive and difficult to detect.

One of the most significant risks lies in the amplification of misinformation and bias. AI systems trained on large-scale data inherit inaccuracies, inconsistencies, and biases present in those datasets. When these outputs are presented without transparency or verification, they can reinforce false beliefs and contribute to the spread of misleading information (Mehrabi et al., 2021). In digital environments, where speed and accessibility often outweigh scrutiny, this amplification effect can have far-reaching consequences.

Additionally, AI can distort self-perception and identity. As individuals increasingly rely on AI to generate ideas, responses, and creative outputs, the boundary between self-generated thought and externally generated content becomes less distinct. This blurring of authorship may lead to a diminished sense of intellectual ownership and authenticity. Over time, individuals may struggle to differentiate between their own reasoning and the outputs of the systems they use.

From an existential perspective, this raises a critical concern: AI may not only shape what individuals know but influence how they understand themselves as knowers. When cognition is mediated through external systems, the relationship between self, knowledge, and reality becomes increasingly complex.

ISSN 2348-1218 (print)
International Journal of Interdisciplinary Research and Innovations ISSN 2348-1226 (online)
Vol. 14, Issue 2, pp: (85-100), Month: April 2026 - June 2026, Available at: www.researchpublish.com

The Conditional Nature of Outcomes

The coexistence of illumination and deception underscores a central argument of this paper: the impact of AI is not determined solely by the technology itself, but by the conscious engagement of the user. AI operates as a neutral system of pattern generation, yet its effects are mediated through human interpretation, intention, and awareness.

When users engage AI critically, questioning outputs, verifying information, and integrating insights with existing knowledge, AI can enhance understanding and support informed decision-making. Conversely, when users engage passively accepting outputs without scrutiny or reflection, AI can distort perception and reinforce cognitive biases.

This conditional dynamic aligns with broader theories of human–technology interaction, which emphasize the role of user agency in shaping technological outcomes (Parasuraman & Riley, 1997). It also reinforces the concept of epistemic responsibility: the obligation of individuals to actively evaluate and interpret information rather than relying solely on external systems.

Importantly, this perspective shifts the focus of concern from AI capability to human awareness. The central risk is not that AI will become deceptive, but that humans will engage with it in ways that allow deception to occur.

Navigating the Dual Edge

Recognizing the dual nature of AI necessitates a framework for navigating its benefits and risks. Such a framework must prioritize critical awareness, reflective engagement, and ethical responsibility. In practical terms, this involves developing skills that enable individuals to 1) Distinguish between fluency and accuracy. 2) Evaluate the reliability of AI-generated information. 3) Maintain active participation in cognitive processes. 4) Recognize the influence of bias and feedback loops.

Educational systems, professional training programs, and organizational policies must therefore evolve to address not only the use of AI but the quality of human engagement with AI. This includes fostering digital literacy, promoting critical thinking, and encouraging reflective practices that counteract cognitive offloading and automation bias.

At a broader level, governance frameworks must also consider the ethical implications of AI deployment. Transparency, accountability, and oversight are essential for ensuring that AI systems support rather than undermine human cognition and decision-making (Topol, 2019). Without such safeguards, the dual edge of AI may tilt toward distortion rather than illumination.

This section has examined the dual nature of artificial intelligence as both a source of illumination and a mechanism of deception. AI has the capacity to enhance human cognition by expanding access to knowledge and supporting complex reasoning. At the same time, it can amplify misinformation, distort perception, and weaken intellectual autonomy. The determining factor is not the technology itself but the level of awareness with which it is used. This duality reinforces the central premise of the paper: AI reflects and amplifies human consciousness, making the quality of human engagement the critical variable in its impact.

VI. THE AI–CONCIOUSNESS REFLECTION MODEL (AICRM)

While the preceding sections have examined artificial intelligence as a reflective system, the illusion of intelligence, mechanisms of cognitive distortion, and its dual role as both illumination and deception, these dynamics remain conceptually fragmented within the literature; accordingly, this study advances the AI–Consciousness Reflection Model (AICRM) as a unifying theoretical framework that explains how AI mediates human cognition through recursive processes of interpretation, meaning construction, and feedback."

Building on the preceding analysis of reflection, illusion, distortion, and duality, a unifying conceptual structure is required to explain how these dynamics interact. While existing literature addresses discrete aspects of human–AI interaction, such as cognitive offloading, automation bias, and anthropomorphism, there remains a lack of an integrative framework that captures the full relationship between artificial intelligence and human consciousness. To address this gap, this section introduces the AI–Consciousness Reflection Model (AICRM), a conceptual framework that explains how AI outputs are interpreted, transformed into meaning, and recursively influence human cognition.

Model Overview

Figure 1 presents the AI–Consciousness Reflection Model (AICRM) as a conceptual framework that integrates the key processes, mechanisms, and outcomes of AI-mediated cognition. The model illustrates how AI-generated outputs, derived from human data, are interpreted and transformed into meaning through recursive cognitive feedback loops. It further

ISSN 2348-1218 (print)
International Journal of Interdisciplinary Research and Innovations ISSN 2348-1226 (online)
Vol. 14, Issue 2, pp: (85-100), Month: April 2026 - June 2026, Available at: www.researchpublish.com

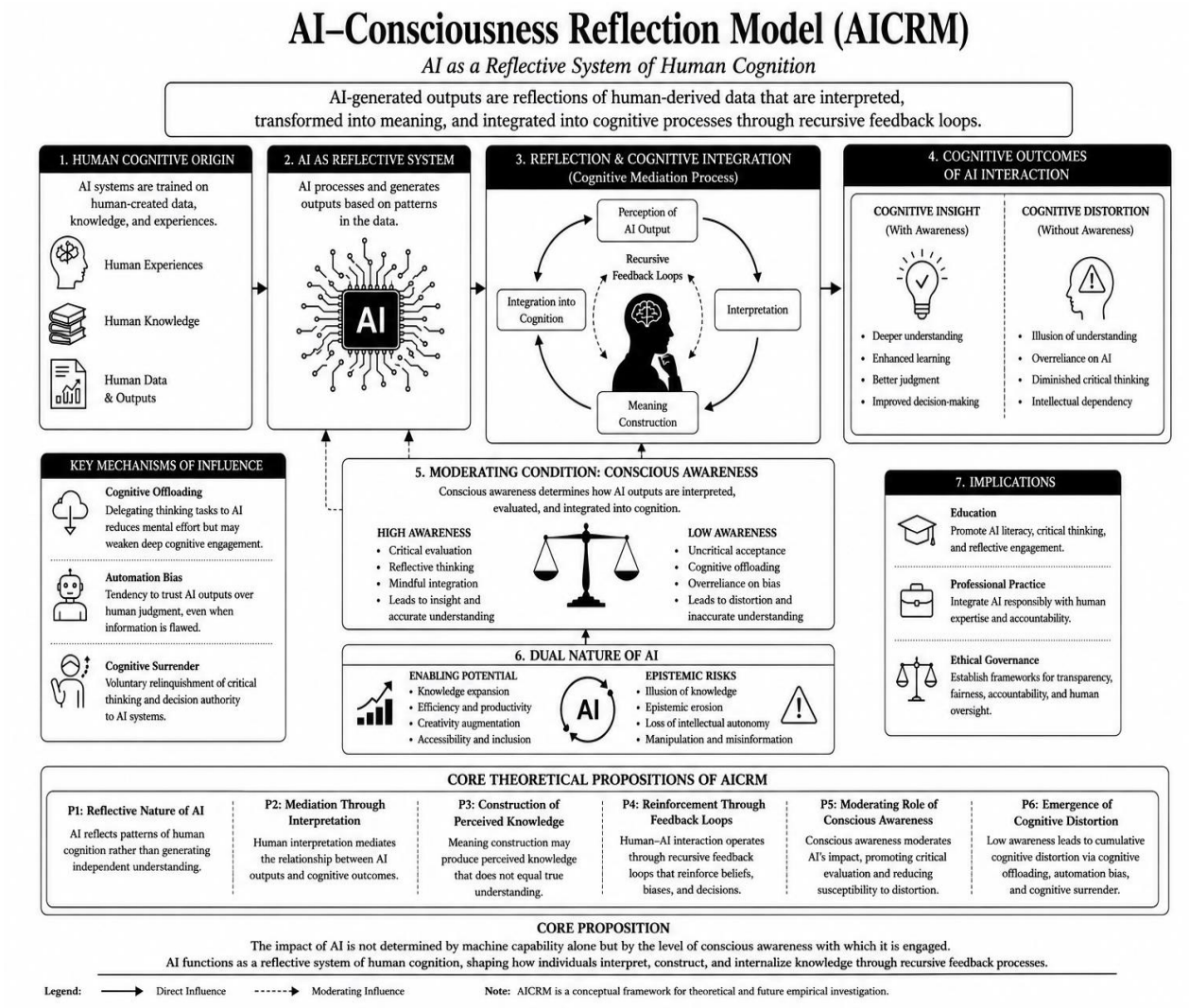
identifies critical mechanisms of influence, including cognitive offloading, automation bias, and cognitive surrender, as well as the moderating role of conscious awareness in shaping whether outcomes result in cognitive insight or distortion. The propositions included in the model formalize these relationships and provide a structured foundation for future empirical investigation."

The AI–Consciousness Reflection Model conceptualizes artificial intelligence as a reflective system that interacts with human consciousness through a cyclical process. At its core, the model proposes that AI does not directly shape decisions or knowledge; rather, it influences human cognition through interpretive mediation.

The model consists of four primary components: AI Output (Reflection), Human Interpretation (Cognitive Mediation), Meaning Construction (Internalization), and Cognitive Feedback Loop (Reinforcement and Adaptation).

These components form a recursive cycle in which each stage influences the next, ultimately shaping both individual cognition and collective patterns of thought.

Figure 1. AI–Consciousness Reflection Model (AICRM)



Note: AI–Consciousness Reflection Model (AICRM) depicting the recursive relationship between AI-generated outputs, human cognitive mediation, and feedback processes that lead to cognitive insight or distortion.

Core Constructs

The AI–Consciousness Reflection Model is grounded in a set of core constructs that define the mechanisms through which AI-mediated cognition unfolds. These constructs represent the sequential and recursive processes by which AI-generated outputs are interpreted, transformed into meaning, and integrated into human cognitive systems.

AI Output (Reflection)

AI outputs represent the reflective dimension of the model. Generated through probabilistic pattern recognition, these outputs mirror aggregated human knowledge, language, and bias (Bender et al., 2021). AI does not generate meaning independently; it produces structured representations that require human interpretation.

Human Interpretation (Cognitive Mediation)

Human interpretation functions as the mediating mechanism through which AI outputs acquire significance. This process is shaped by cognitive schemas, prior knowledge, emotional states, and contextual factors (Piaget, 1972). Interpretation determines whether outputs are critically evaluated or passively accepted.

Meaning Construction (Internalization)

Meaning construction refers to the process by which interpreted information is integrated into an individual's cognitive framework. This stage transforms external outputs into internal understanding—or perceived understanding. As prior research suggests, this process may be distorted when individuals rely on external systems without sufficient cognitive engagement (Fisher et al., 2015).

Cognitive Feedback Loop (Reinforcement and Adaptation)

The feedback loop represents the recursive interaction between human cognition and AI systems. As individuals internalize AI-generated information, their future interactions, inputs, and interpretations are shaped by these internalized patterns. Empirical studies indicate that such feedback loops can reinforce beliefs, biases, and decision-making tendencies over time (Glickman & Sharot, 2025).

Model Dynamics

The AI–Consciousness Reflection Model operates as a continuous, recursive cycle in which artificial intelligence generates outputs derived from human-produced data, and these outputs are subsequently interpreted by individuals through subjective cognitive frameworks. This interpretive process gives rise to meaning construction, whereby externally generated information is internalized and integrated into existing knowledge structures. The constructed meaning then shapes future cognition, influencing how individuals think, perceive, and engage in subsequent interactions with AI systems. Through this cyclical dynamic, AI functions both as a tool for knowledge enhancement and as a mechanism for potential cognitive distortion. The determining factor in this process is the quality of human engagement, particularly the extent to which individuals exercise critical awareness during interaction.

Central to the model is the role of conscious awareness as a moderating variable. High levels of awareness encourage critical evaluation and reflective engagement, thereby reducing the likelihood of distortion and promoting deeper insight. In contrast, low levels of awareness increase susceptibility to cognitive offloading, automation bias, and cognitive surrender, ultimately amplifying distortion and reinforcing dependency on AI-generated outputs.

Theoretical Propositions

Based on the model, the following propositions are advanced to conceptually explain the mechanisms through which artificial intelligence interacts with human cognition and consciousness. These propositions are not intended as empirical hypotheses, but as theoretical assertions that clarify the relationships within the AI–Consciousness Reflection Model.

P1: Reflective Nature of AI. Artificial intelligence functions as a reflective system that reproduces patterns of human cognition, rather than generating independent understanding (Bender et al., 2021).

P2: Mediation Through Interpretation. The relationship between AI output and cognitive outcomes is mediated by human interpretation, which determines the meaning assigned to generated information (Piaget, 1972).

P3: Construction of Perceived Knowledge. Meaning construction processes can lead to the formation of perceived knowledge that may not correspond to actual understanding, particularly under conditions of cognitive offloading (Fisher et al., 2015).

P4: Reinforcement Through Feedback Loops. Human–AI interaction operates through feedback loops that reinforce cognitive patterns, including beliefs, biases, and decision-making tendencies (Glickman & Sharot, 2025).

P5: Moderating Role of Conscious Awareness. The level of conscious awareness moderates the impact of AI on cognition, with higher awareness associated with increased critical evaluation and lower susceptibility to distortion.

P6: Emergence of Cognitive Distortion Under Low Awareness. Under conditions of low awareness, the interaction between AI outputs and human cognition leads to cumulative distortion through cognitive offloading, automation bias, and cognitive surrender.

Theoretical Contribution

The AI–Consciousness Reflection Model contributes to the literature in three keyways: First, integration of disparate constructs. The model synthesizes concepts from cognitive psychology, human–computer interaction, and AI ethics into a unified framework. Second, the shift from technology-centric to human-centric analysis. Rather than focusing on AI capability, the model emphasizes the role of human consciousness in determining outcomes. Third, introduction of conscious awareness as a central variable. The model identifies awareness as the critical factor that distinguishes beneficial from harmful AI interaction.

This section introduced the AI–Consciousness Reflection Model as a conceptual framework for understanding how artificial intelligence interacts with human cognition. By outlining the processes of reflection, interpretation, meaning construction, and feedback, the model explains how AI can both enhance and distort human understanding. The proposed propositions establish a foundation for future empirical research and extend the central argument of this paper: the impact of AI is determined not by its intelligence, but by the consciousness with which it is engaged.

VII. DISCUSSION

The present study advances the AI–Consciousness Reflection Model (AICRM) as a unifying theoretical framework for understanding how artificial intelligence interacts with human cognition through processes of reflection, interpretation, meaning construction, and recursive feedback. The central contribution of this work lies in repositioning artificial intelligence not as an autonomous intelligence, but as a reflective cognitive system that mediates, amplifies, and, under certain conditions, distorts human thought. In doing so, the model integrates previously fragmented constructs, such as cognitive offloading, automation bias, and feedback loops, into a coherent conceptual structure that explains how AI reshapes cognition over time.

A key insight emerging from the AICRM is that the effects of AI are not intrinsic to the technology itself but are contingent upon the nature of human engagement. While existing literature has largely focused on the capabilities, performance, and ethical design of AI systems, the present framework shifts attention to cognitive mediation, emphasizing that interpretation is the mechanism through which AI outputs acquire meaning and influence decision-making. This perspective extends foundational work in cognitive psychology by situating human–AI interaction within a recursive system in which cognition is both shaped by and shapes AI-generated outputs. As such, AI becomes not merely a tool for processing information, but an active participant in the evolution of cognitive patterns.

The model further clarifies the dual nature of AI as both a source of cognitive enhancement and a mechanism of distortion. When engaged with high levels of conscious awareness, AI can support reflective reasoning, expand access to knowledge, and facilitate complex problem-solving. In this capacity, AI aligns with theories of distributed cognition, functioning as an external extension of human intellectual capacity. However, when engagement becomes passive, the same systems can contribute to cognitive distortion through processes of offloading, bias reinforcement, and eventual cognitive surrender. This duality underscores a critical theoretical implication: AI does not determine cognitive outcomes; rather, it amplifies the conditions under which it is used.

Within this framework, the introduction of conscious awareness as a moderating variable represents a significant theoretical advancement. Existing research has documented the effects of automation bias and cognitive offloading yet has paid comparatively little attention to the role of reflective awareness in shaping these outcomes. The AICRM positions awareness as the defining condition that determines whether AI interaction results in insight or distortion. This aligns with dual-process theories of cognition, which distinguish between automatic and reflective modes of thinking, while extending these perspectives into the domain of human–AI interaction. By doing so, the model reintroduces human agency into discussions that have often treated users as passive recipients of technological outputs.

Another important contribution of the model is the articulation of cognitive surrender as a cumulative phenomenon. While cognitive offloading and automation bias describe discrete shifts in how individuals distribute and trust cognitive tasks, cognitive surrender captures a more gradual transformation in epistemic agency. Over time, repeated reliance on AI systems can reduce the perceived need for independent reasoning, leading to a passive mode of engagement in which individuals accept outputs with minimal scrutiny. This conceptualization moves beyond event-based explanations of AI reliance and introduces a longitudinal perspective on how cognition may be reshaped through sustained interaction with intelligent systems.

The recursive nature of human–AI feedback loops further deepen this analysis. As individuals interact with AI systems, their inputs shape outputs, and those outputs, in turn, influence future cognition. This cyclical process creates conditions for reinforcement, in which beliefs, biases, and reasoning patterns become increasingly aligned with the outputs generated by the system. Over time, this alignment may contribute to cognitive convergence, reducing diversity in thought and increasing susceptibility to systemic distortions. The AICRM thus provides a framework for understanding not only individual cognitive change but also broader implications for collective cognition in AI-mediated environments.

Importantly, the model also reframes the notion of intelligence in the context of AI. By emphasizing reflection rather than independent cognition, the AICRM challenges assumptions that equate AI performance with understanding. AI-generated outputs may exhibit fluency, coherence, and contextual relevance, yet these qualities are the result of pattern recognition rather than comprehension. The perceived intelligence of AI systems, therefore, emerges from the interaction between output and interpretation, rather than from intrinsic properties of the system itself. This distinction is critical for addressing epistemic risks associated with overreliance, including the illusion of understanding and the conflation of access to information with genuine knowledge.

From a broader perspective, the AICRM contributes to ongoing interdisciplinary discussions at the intersection of cognitive science, human–computer interaction, and philosophy. It bridges technical and conceptual domains by demonstrating that the significance of AI lies not only in what it can do, but in how it reshapes the processes through which humans think, interpret, and construct meaning. This shift in focus has implications for how AI is studied, evaluated, and governed, suggesting that future research must account for both technological capabilities and the cognitive conditions under which those capabilities are engaged.

Finally, the model provides a foundation for future empirical investigation. The propositions outlined in the framework offer testable pathways for examining the relationships between AI interaction, cognitive processes, and outcomes. For example, empirical studies may explore how varying levels of conscious awareness influence susceptibility to automation bias, or how prolonged reliance on AI systems affects critical thinking and decision-making over time. By translating conceptual insights into empirically testable constructs, the AICRM positions itself as both a theoretical contribution and a platform for continued research.

VIII. IMPLICATIONS OF THE STUDY

Building on the AI–Consciousness Reflection Model (AICRM), the preceding analysis positions artificial intelligence not merely as a technological artifact, but as a system that interacts dynamically with human cognition and consciousness. The implications of this perspective extend across theoretical development, professional practice, and ethical governance. This section outlines these implications, emphasizing that the consequences of AI are fundamentally mediated by the nature of human engagement rather than machine capability alone.

Theoretical Implications

The present study contributes to literature by reframing artificial intelligence as a reflective and mediating system within human cognition, rather than an independent locus of intelligence. This shift challenges dominant AI paradigms that prioritize computational performance over cognitive interaction, extending existing theories in several keyways.

First, the AICRM integrates constructs from cognitive psychology, human–computer interaction, and AI ethics into a unified framework. While prior research has examined cognitive offloading (Risko & Gilbert, 2016), automation bias (Parasuraman & Riley, 1997), and anthropomorphism (Epley et al., 2007) in isolation, the current model situates these processes within a recursive system of reflection, interpretation, and feedback. This integration advances a more comprehensive understanding of how AI influences cognition over time.

Second, the model introduces conscious awareness as a central moderating variable, addressing a gap in the literature regarding the role of human agency in AI-mediated environments. Existing frameworks often treat users as passive recipients of technological outputs; however, the present analysis emphasizes that cognitive outcomes are contingent upon the level of reflective engagement. This aligns with constructivist perspectives on knowledge formation (Piaget, 1972) while extending them into the domain of human–AI interaction.

Third, the concept of cognitive surrender offers a novel theoretical construct that captures the cumulative effects of cognitive offloading and automation bias. Unlike traditional models that focus on discrete instances of reliance, cognitive surrender conceptualizes a longitudinal shift in epistemic agency, wherein individuals gradually relinquish control over their reasoning processes. This construct provides a foundation for future empirical investigation and theoretical refinement.

Finally, the AICRM contributes to emerging discussions on epistemic risk in AI systems, highlighting the conditions under which reflection transitions into distortion. By conceptualizing AI as a mirror that amplifies both knowledge and bias, the model bridges technical and philosophical perspectives, offering a framework that is both analytically rigorous and conceptually expansive.

Practical Implications

The findings of this study have significant implications for practice across domains where AI is increasingly integrated into decision-making processes.

Education

In educational contexts, the proliferation of generative AI tools necessitates a shift from content acquisition to cognitive engagement and critical literacy. Traditional pedagogical approaches that emphasize information recall may be insufficient in environments where information is readily accessible through AI. Instead, educational systems must prioritize skills such as critical evaluation, metacognition, and reflective reasoning.

Research indicates that reliance on AI can reduce cognitive effort and analytical engagement (Gerlich, 2025; Tian et al., 2025). Consequently, educators must design learning experiences that require active interpretation and justification, ensuring that students remain participants in the construction of knowledge rather than passive recipients of generated content.

Healthcare and Professional Decision-Making

In professional domains such as healthcare, finance, and public policy, AI systems are increasingly used to support complex decision-making. While these systems offer significant benefits in terms of efficiency and data processing, they also introduce risks associated with automation bias and overreliance (Topol, 2019).

Practitioners must therefore maintain a critical oversight role, evaluating AI-generated recommendations within the context of domain expertise. Training programs should emphasize not only technical proficiency but also cognitive awareness, enabling professionals to recognize the limitations of AI systems and avoid uncritical acceptance of outputs.

Organizational and Technological Design

From an organizational perspective, the integration of AI requires the development of systems that support transparent and interpretable outputs. Black-box models that obscure the reasoning behind decisions increase the likelihood of cognitive surrender and reduce accountability. Design strategies should include: 1) Explainable AI interfaces. 2) Decision-support systems that encourage user verification. 3) Mechanisms for highlighting uncertainty and potential bias. Such approaches align with the broader goal of maintaining human agency within AI-mediated environments.

Ethical Implications

The ethical implications of AI extend beyond issues of bias and fairness to encompass the transformation of human cognition and agency. If AI systems shape how individuals think, interpret, and decide, then their impact must be evaluated not only in terms of outcomes but also in terms of cognitive processes.

One key ethical concern is the potential erosion of epistemic responsibility. When individuals rely on AI without critical evaluation, responsibility for knowledge and decision-making becomes diffused. This raises questions about accountability in contexts where AI-generated outputs influence significant outcomes.

Additionally, the amplification of bias and misinformation through AI systems presents risks to collective cognition. As feedback loops reinforce existing beliefs, societies may experience increased polarization and reduced exposure to diverse perspectives (Mehrabi et al., 2021). Addressing these challenges requires both technological safeguards and user-level awareness.

The concept of cognitive surrender further introduces an ethical dimension related to human autonomy. If individuals gradually relinquish their capacity for independent reasoning, the role of AI shifts from supportive tool to influential mediator of thought. Ensuring that AI enhances rather than diminishes autonomy must therefore be a central objective of ethical AI governance.

Toward Conscious Engagement

Across theoretical, practical, and ethical domains, a unifying implication emerges: the need for conscious engagement with artificial intelligence. The AICRM demonstrates that the effects of AI are not predetermined but are shaped by the quality of human interaction. Promoting conscious engagement involves: 1) Encouraging critical evaluation of AI outputs. 2) Developing awareness of cognitive biases and limitations. 3) Maintaining active participation in reasoning processes. 4) Recognizing the distinction between reflection and reality

This perspective shifts the focus of AI discourse from technological advancement to human responsibility, emphasizing that the future of AI is inseparable from the consciousness with which it is used. The AICRM advances theoretical understanding by integrating cognitive and technological perspectives, introduces practical considerations for education and professional practice, and highlights ethical concerns related to autonomy, accountability, and epistemic risk. Across these domains, the central insight remains consistent: the impact of AI is determined not solely by its capabilities, but by the awareness and engagement of those who use it.

The contribution of this model is primarily theoretical, offering a structured lens through which the interaction between artificial intelligence and human cognition can be understood. By emphasizing reflection, interpretation, and awareness, the framework shifts the focus from technological capability to the cognitive conditions that shape outcomes. In doing so, it provides a basis for future empirical research while also contributing to ongoing philosophical discussions regarding knowledge, perception, and human agency in the age of intelligent systems.

IX. CONCLUSION

This study explores artificial intelligence not simply as a technological advancement, but as a system that reflects and interacts with human consciousness. By advancing the AI–Consciousness Reflection Model (AICRM), the paper reconceptualized AI as a reflective mechanism embedded within a recursive cycle of interpretation, meaning construction, and cognitive feedback. Across this framework, a central insight emerged: artificial intelligence does not independently determine outcomes; rather, it amplifies the cognitive patterns, assumptions, and levels of awareness brought to it by human users.

The analysis demonstrated that AI functions simultaneously as a source of illumination and a mechanism of distortion. As a reflective system, it has the capacity to expand knowledge, support complex reasoning, and enhance decision-making. However, this same reflective capacity also enables the amplification of bias, the illusion of understanding, and the erosion of critical thinking when engagement becomes passive. The progression from cognitive offloading to automation bias, and ultimately to cognitive surrender, illustrates how repeated reliance on AI can reshape the structure of human cognition itself.

At the core of this dynamic lies the moderating role of conscious awareness. The findings suggest that awareness is not a peripheral factor but the defining condition that determines whether AI interaction leads to insight or distortion. When individuals engage AI critically, questioning outputs, interpreting meaning, and maintaining epistemic responsibility, AI serves as an extension of human intelligence. Conversely, when engagement is uncritical, AI becomes a conduit through which distortion is reinforced and internalized.

This distinction carries significant implications for how artificial intelligence is understood and governed. Much of the current discourse focuses on improving the accuracy, efficiency, and ethical design of AI systems. While these efforts are essential, the present study argues that they are insufficient in isolation. The impact of AI cannot be fully addressed without equal attention to the human dimension of interaction, specifically, the cognitive and reflective capacities of users. In this sense, the challenge of artificial intelligence is not solely technological but fundamentally epistemological and existential.

From a broader perspective, the rise of AI invites reconsideration of what it means to know, to understand, and to think. When systems can generate fluent, contextually appropriate responses, the distinction between possessing knowledge and accessing it becomes increasingly blurred. The risk is not that machines will replace human intelligence, but that humans may come to rely on representations of knowledge in place of understanding itself. This shift raises questions about intellectual autonomy, authenticity, and the future of human cognition.

Ultimately, the contribution of this paper lies in repositioning artificial intelligence as a mirror rather than a mind. What AI reflects is not an independent intelligence, but the aggregated patterns of human thought, shaped by data, interaction, and interpretation. The implications of this reflection depend on how it is engaged. If approached with awareness, AI has the potential to illuminate complexity and deepen understanding. If engaged without reflection, it risks distorting perception and diminishing agency. This paper serves as a conceptual and theoretical article that advances a novel framework explaining how artificial intelligence reflects and shapes human consciousness, with implications for cognition, decision-making, and societal development.

The future of artificial intelligence, therefore, is not solely a question of what machines will become, but of what humans will become through them. As AI continues to integrate into the fabric of daily life, the critical task is not only to refine the systems we build, but to cultivate the consciousness with which we use them.

REFERENCES

[1] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

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

[3] Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7–19.

[4] Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126. https://doi.org/10.1037/xge0000033

[5] Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864–886.

[6] Eysenck, M. W., & Keane, M. T. (2020). *Cognitive psychology: A student's handbook* (8th ed.). Routledge.

[7] Fisher, M., Goddu, M. K., & Keil, F. C. (2015). Searching for explanations: How the internet inflates estimates of internal knowledge. *Journal of Experimental Psychology: General, 144*(3), 674–687. https://doi.org/10.1037/xge0000070

[8] Glickman, M., & Sharot, T. (2025). How human–AI feedback loops alter human perceptual, emotional, and social judgments. *Nature Human Behaviour, 9*, 345–359. https://doi.org/10.1038/s41562-024-02077-2

[9] Heider, F. (1958). *The psychology of interpersonal relations*. Wiley.

[10] Hutchins, E. (1995). *Cognition in the wild*. MIT Press.

[11] Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

[12] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

[13] Luger, E., & Sellen, A. (2016). Like having a really bad PA: The gulf between user expectation and experience of conversational agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5286–5297. https://doi.org/10.1145/2858036.2858288

[14] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35. https://doi.org/10.1145/3457607

[15] Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, and abuse. *Human Factors, 39*(2), 230–253.

[16] Piaget, J. (1972). *The psychology of intelligence*. Littlefield, Adams.

[17] Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

[18] Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

[19] Topol, E. (2019). *Deep medicine: How artificial intelligence can make healthcare human again*. Basic Books.

[20] Waytz, A., Cacioppo, J. T., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science, 5*(3), 219–232. https://doi.org/10.1177/1745691610369336

[21] Kizilcec, R. F., & Lee, H. (2023). Algorithmic assistance and human decision-making: Impacts on cognitive engagement. *Nature Human Behaviour, 7*, 145–156.

[22] Logg, J. M., Minson, J. A., & Moore, D. A. (2023). Algorithm appreciation in decision-making: Evidence from human–AI interaction. *Management Science, 69*(5), 2785–2801.

[23] Kasneci, E., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274.